Universidade de Lisboa

Faculdade de Ciências

Departamento de Biologia Animal

Family Genetics of Paget's Disease of Bone

Patrícia Alexandra Silva Santos

Dissertação

Mestrado em Biologia Humana e Ambiente

2014

Universidade de Lisboa

Faculdade de Ciências

Departamento de Biologia Animal

Family Genetics of Paget's Disease of Bone

Patrícia Alexandra Silva Santos

Dissertação

Mestrado em Biologia Humana e Ambiente

Dissertação orientada por:

Professora Doutora Sofia A. Oliveira (IMM/FMUL)

Professora Doutora Deodália Dias (DBA/FCUL)

2014

Acknowledgements

First, I would like to thank my supervisor Doctor Sofia Oliveira for giving me the opportunity to perform my master thesis project in her lab, for all the help and support in driving forward this project the best way possible.

I also want to thank to Doctor Inês Sousa for her supervision, assistance, support, friendship and more important all the patience in teaching me and help me achieve more each time. To MSc. Vânia Francisco for all the hours spent teaching me, for all the support, friendship and also a lot of patience. Their contributions were very important to me. You have made this experience one of the greatest, an opportunity to grow up both academic and at a personal level.

A special thanks to Doctor Daniel Sobral and his team (Renato Alves, Paulo Almeida and Isabel Marques) for giving me the opportunity to carry out all the bioinformatics analysis in their unit and especially for all the support and time spent teaching me.

To Doctor José Vaz Patto and Doctor Filipe Barcelos, for their availability and help over the past one year. And also to the two families who agreed to participate in my project and made it a reality.

To Susana Ladeiro, from the Genomics Unit (*Instituto Gulbenkian de Ciência*), that made the validation of the novel variants identified.

To Joana Xavier for previous patient blood collection and to Nurse Conceição Afonso for collecting patient blood samples in *Alentejo*.

To all my friends, especially to Ana Rebelo, Nádia Rocha, Sara Bonucci, Rodrigo Godinho and Tiago Ferreira, for always being there for me, all the support, friendship and the hours spent listening to my "stresses".

To Fernanda Estevez for all the support and friendship.

Lastly, a special thanks to my family who always supported me in every decision I made and for making the last year possible. Without you I could not have achieved half of what I have already.

Contents

Note: This thesis follow the citation style of the Nature Genetics journal.

List of tables

List of figures

List of abbreviations

1000 GP	1000 Genomes Project
AcNa	Sodium acetate
AD	Depth Allele by sample
ALT	Alternate
ASP	Affected Sibling Pairs
BAM	Binary Alignment Files
BLAT	Blast-Like Alignment Tool
BGI	Beijing Genomics Institute
bp	Base pairs
BWA	Burrows-Wheeler Aligner
CDS	Coding DNA Sequence
cM	Centimorgan
dbSNP	Single Nucleotide Polymorphism Database
ddATP	Dideoxyadenosine triphosphate
ddCTP	Dideoxycytidine triphosphate
ddGTP	Dideoxyguanosine triphosphate
ddNTPs	Dideoxyribonucleotide triphosphates
ddTTP	Dideoxythymidine triphosphate
DNA	Deoxyribonucleic acid
dNTPs	Deoxyribonucleotide triphosphates
DP	Read DePth
DSSP	Dictionary of Protein Secondary Structure
EDTA	Ethylenediaminetetraacetic Acid
eoPDB	Early-onset PDB
ESH	Expansile Skeletal Hyperphosphatasia
FEO	Familial Expansile Osteolysis
FS	Fisher Strand
g	Relative centrifugal force
GATK	Genome Analysis Toolkit
GQ	Genotype Quality
GWAS	Genome-Wide Association Study
IBMPFD	Inclusion Body Myopathy combined with PDB and Frontotemporal Dementia
IGC	*Instituto Gulbenkian de Ciência*
IMM	*Instituto de Medicinal Molecular*
InDel	Insertion or Deletion of bases in the DNA
IPR	*Instituto Português de Reumatologia*
JPD	Juvenile PDB
LD	Linkage Disequilibrium
LOD	Logarithm of the odds

MAF	Minor Allele Frequency
MAPQ	MAPping Qualiy
MIGS	Minimum Information about a Genome Sequence
MQRankSum	Mapping Quality Rank Sum Test
NFD	Non-Frameshift Deletion
NFκB	Nuclear factor kappa B
NGS	Next-Generation Sequencing
NSV	Non-Synonymous Variants
NTX	N-terminal telopeptide excretion
OCL	Osteoclasts
OPTN	*Optineurin* gene
PCR	Polymerase Chain Reaction
PDB	Paget's Disease of Bone
Q	Base quality value
QC	Quality Control
QD	Quality by Depth
QUAL	Quality score for SNVs and InDels
RANK	Receptor Activator of Nuclear Factor κB
RBC	Red Blood Cells
REF	Reference
RNA	Ribonucleic acid
RPM	Revolutions Per Minute
SAM	Sequence Alignment/Map
SIFT	Sorting Tolerant From Intolerant
SNP	Single-Nucleotide Polymorphism
SNV	Single Nucleotide Variant
STR	Short Tandem Repeat
SQSTM1	*Sequestosome 1* gene
TAE	Tris-Acetate-EDTA
TGF-β	Tumor Growth Factor-β
Tm	Melting temperature
TNF	Tumor Necrosis Factor
UTR	Untranslated Region
UV	Ultra Violet light
VCP	*Valosin-Containing Protein* gene
VDR	*Vitamin D Receptor* gene
VEP	Variant Effect Predictor
WES	Whole-Exome Sequencing
WGS	Whole-Genome Sequencing

Abstract

Paget's disease of bone (PDB) is a systemic disease characterized by increased bone resorption and formation, causing gradual destruction of parts of the skeleton and subsequent reconstruction of a more fragile bone. PDB has an overall incidence of 2% in the population over 55 years. PDB is a complex disease with multiple genes implicated in its pathogenesis, but in its monogenic form, only one gene (*SQSTM1*) has been linked to PDB.

To identify novel genes causing familial PDB, we performed whole exome sequencing (WES) in six individuals from a Portuguese multiplex family composed of five PDB cases, two unaffected individuals and one individual with unclear diagnosis. Given the uncertain diagnosis for one family member, we conducted two analyses: model 1, in which this individual is considered affected and model 2 where he is unaffected. DNA was captured using the SureSelect Target Enrichment System kit and sequenced using Hiseq2000 (Illumina's Solexa). We identified three variants (c.C4786T (*KIAA1875*), c.C53T (*NLRC3*) and c.T566C (*SRL*)) in model 1 and one variant (c.G180A (*SERINC2*)) in model 2 that were present in all affected and absent from the unaffected in next-generation sequencing (NGS) data. Validation of these mutations by Sanger sequencing in all family members revealed that all model 1 mutations were present in all individuals, while the model 2 mutation was present in all family members except the individual with unclear diagnosis. None of these variants were present in a second Portuguese PDB multiplex family.

In conclusion, our findings support the notion that bioinformatics analyses of NGS data is a process requiring optimization. We found four novel variants which may cause PDB in this family with an autosomal dominant pattern of inheritance and incomplete penetrance. Further studies in other PDB families are warranted to determine the pathogenic potential of these genes/variants.

Sumário

Os ossos são um tecido importante do corpo humano com diversas funções, tais como, a protecção dos órgãos, armazenamento de minerais e reservatório para diversas células. Três células essências contribuem para que o tecido ósseo seja continuamente remodelado, preservando a homeostase mineral. Essas células são os osteoclastos, responsáveis pela reabsorção óssea, os osteoblastos, responsáveis pela formação óssea, e os osteócitos, responsáveis pela manutenção da integridade da matriz óssea. Em condições normais, a interacção entre as três células mantém a remodelação óssea equilibrada. No entanto, quando há um desequilíbrio na remodelação óssea podem-se desenvolver doenças, tais como a doença óssea de Paget (DOP).

A DOP é uma doença sistémica em que a taxa de reabsorção e formação óssea estão aumentadas, causando uma destruição gradual de partes do osso, e uma consequente reconstrução de um osso mais frágil e desorganizado. A DOP é uma doença complexa com uma incidência de aproximadamente 2% na população com mais de 55 anos. 85% dos indivíduos afectados não manifesta qualquer sintoma, recebendo o diagnóstico para a DOP quando faz exames médicos de rotina. A percentagem de pacientes com sintomas manifesta dores, osteoartrite, fracturas, surdez, entre outros.

Estudos sugerem que a doença surgiu no Reino Unido tendo-se depois dispersado para outras partes do mundo, provavelmente devido a fenómenos migratórios. A DOP é assim mais comum no Reino Unido, Austrália, Nova Zelândia, África do Sul, Ásia e em Portugal onde se verifica um foco de indivíduos afectados no Alentejo, mas não se conhece ainda uma causa para esse facto.

A etiologia da DOP para a maioria dos casos é desconhecida. Pensa-se que a combinação de variantes comuns e/ou raras em conjunto com factores ambientais que despoletam a manifestação da doença. Dos factores ambientais estudados, o mais defendido como desencadeador da DOP é a infecção pelo paramamyxovirus, tendo-se observado no núcleo e citosol dos osteoclastos partículas parecidas com nucleocapsídeos de paramyxovirus. Os osteoclastos de doentes com DOP manifestam um fenótipo diferentes dos indivíduos não afectados, tendo mais núcleos por célula, resistência à apoptose e reabsorção óssea aumentada. No entanto, os factores ambientais requerem a co-expressão de factores genéticos para o desenvolvimento da doença.

Até ao momento, o *SQSTM1* é o único gene identificado responsável pelo desenvolvimento da DOP familiar. No entanto, estudos de linkage e associação apontam para um risco aumentado para o desenvolvimento de DOP idiopática em indivíduos com polimorfismos nos genes *CSF1*, *OPTN*, *TNFRSF11A*, *TM7SF4*, *NUP205*, *RIN3*, *PML* e *GOLGA6A*. Contudo, são necessários mais estudos para identificar outras variantes causais que explicam a restante variabilidade genética.

O nosso objectivo é identificar o(s) gene(s) que causa a DOP numa família Portuguesa com vários indivíduos afectados com DOP oriunda do Alentejo. Para tal, foi solicitado à Beijing Genomics Institute (BGI) Hong Kong que fizesse a sequenciação do exoma (whole-exome sequencing [WES]) dos seis familiares usando next-generation sequencing (NGS). NGS é uma técnica de sequenciação que produz dezenas de fragmentos de DNA (entre 100 a 500 pares de base) num curto período de tempo, que tem vindo a ser aplicada no estudo de diversas doenças monogénicas e complexas. WES é um método robusto para a identificação de variantes raras associadas a doenças complexas visto que o exoma (1-2% do genoma que codifica proteínas) representa uma grande parte do genoma onde é possível identificar variantes que produzem efeitos funcionais. O DNA foi capturado usando o kit SureSelect Target Enrichment System e sequenciado usando a plataforma Hiseq2000 (Illumina's Solexa).

Após verificarmos que a análise bioinformática da BGI continha diversos erros realiza-mos uma nova análise seguindo a *pipeline* do Broad Institute (*Genome Analysis Toolkit (GATK) Best Practices workflow*) optimizada para DNA humano, utilizando algumas ferramentas complementares. A *pipeline* encontra-se dividida em três fases, o processamento dos dados (alinhamento dos fragmentos de DNA com a sequência referência e remoção dos duplicados), a detecção das variantes, e por último a anotação e análise das funções das variantes usando ferramentas *in silico* (SIFT e PolyPhen-2). Por fim, os dados foram divididos em polimorfismos nucleótidicos simples (SNPs) e inserção e delecção de nucleótidos (InDels), contendo as variantes encontradas para cada indivíduo na região codificante. Para a análise dos SNPs, dividiram-se as variantes em não-sinónimas, sinónimas, ganho de um codão stop (*stop-gain*) e perda de um codão stop (*stop-loss*). Para a análise dos InDels dividiram-se as variantes em *frameshift insertion/deletion*, *non-frameshift insertion/deletion*, *frameshift/non-frameshift block substitution*, ganho de um codão stop (*stop-gain*) e perda de um codão stop (*stop-loss*). Analisámos ainda as variantes presentes nas regiões regulatórias mas não as validámos.

Assumindo que a mutação que causa a DOP nesta família é privada e nova excluímos todas as variantes presentes em bases de dados públicas (dbSNP e 1000 Genomes Project). Como não foi possível obter até à conclusão desta tese o diagnóstico final de um dos familiares, criámos dois modelos de análise distintos, em que para o modelo 1 o indivíduo é considerado afectado e no modelo 2 é considerado não afectado.

Pretendemos assim identificar as variantes na região codificante que estão presentes nos indivíduos afectados e ausentes no(s) indivíduo(s) não afectado(s). Validar as variantes identificadas por sequenciação de Sanger não só nos indivíduos sequenciados por WES, mas também nos restantes membros da família e numa nova família oriunda do Alentejo e ver a sua segregação. Estamos interessados em analisar variantes novas (ou seja, que não estão descritas em nenhuma base de dados) visto já existerem diversos estudos em variantes comuns que apenas explicam uma parte da variabilidade genética para a doença.

Três novas variantes (c.C4786T (*KIAA1875*), c.C53T in (*NLRC3*) e c.T566C (*SRL*)) foram identificadas utilizando o modelo 1. No modelo 2 identificámos duas novas variantes, c.G180A (*SERINC2*) e uma delecção no *PLEKHG5* (c.2163_2168del). As funções destes genes parecem estar associadas com o metabolismo ósseo, no entanto são necessários mais estudos para confirmar esta associação.

A validação destas variantes por sequenciação de Sanger revelou que as três variantes identificadas para o modelo 1 (c.C4786T, c.C53T, and c.T566C) estavam presentes em todos os indivíduos afectados da família 1. No entanto, estas variantes também estavam presentes nos dois indivíduos controlo. Para o modelo 2, verificou-se que a variante c.2163_2168del insere-se numa posição diferente do genoma quando comparado com os resultados de WES, estando já reportada em bases de dados (estando presente em todos os indivíduos afectados e nos dois controlos da família 1, de acordo com a sequenciação de Sanger). Para o modelo 2, a variante c.G180A está presente em todos os indivíduos afectados e nos dois controlos, no entanto está ausente do individuo controlo adicional (incluído apenas no modelo 2). Numa segunda família multiplex Portuguesa com DOP nenhuma destas mutações foi detectada em nenhum membro da família.

Podemos concluir que nenhuma destas variantes tem uma segregação perfeita do tipo autossómico dominante com penetrância completa, o que está de acordo com o descrito na literatura. O potencial patogénico destas mutações não pode ser excluído se tivermos em conta que alguns indivíduos da família podem não ter sido ainda

diagnosticados clinicamente por não apresentarem sintomas ou por ainda não terem idade para manifestar a DOP.

Estudos adicionais devem incluir variantes com uma frequência do alelo menor inferior a 5% e descritas no dbSNP ou 1000 GP. Há várias variantes localizadas em zonas regulatórias que apontam para uma possível associação com a PDB, sendo necessário mais estudos que incluam variantes que se encontram nestas regiões. O número de indivíduos e famílias no estudo devería de ser maior, de modo a melhorar a identificação das variantes causais, e os indivíduos controlo a estudar deveriam de ter uma idade superior à de risco (acima dos 50 anos).

A identificação de novas variantes genéticas associadas à DOP pode ajudar a compreender melhor as vias celulares envolvidas na sua patogénese. Deste modo poderão, eventualmente, advir novas terapias mais eficazes e porventura preventivas.

Keywords

Paget's disease of Bone (PDB)

Complex disorder

KIAA1875, *NLRC3*, *SRL*, *SERINC2*

Sanger sequencing

Whole-exome sequencing (WES)

Bioinformatics analysis

Palavras-chave

Doença Óssea de Paget (DOP)

Doença complexa

KIAA1875, *NLRC3*, *SRL*, *SERINC2*

Sequenciação de Sanger

Sequenciação do exoma (WES)

Análise bioinformática

1. Introduction

1.1 Paget's disease of bone

The human body is supported and shaped by the bones, which are a complex and highly organized tissue with several functions: organ protection, anchor points for muscles, tendons and ligaments, storage for minerals, and a reservoir for a broad range of cells (such as stem cells of the mesenchymal and hematopoietic cell lineages)[1,2]. Three essential cells, osteoclasts (OCL, bone-resorbing), osteoblasts (bone-forming), and osteocytes (maintenance of the bone matrix health), all contribute for the bone tissue to be continuously remodeled preserving mineral homeostasis, and to maintain its robustness. The damaged or old bone is resorbed by OCL during remodeling, and then, osteoblasts migrate to this resorbed area to form new bone. Once the new bone is formed, osteoblasts are entrapped in the newly mineralized bone matrix and called osteocytes. The latter are able to sense the fluctuations in mechanical load and in hormone levels, amongst others, forming a large signaling network to communicate with each other, lining cells on the bone surface and with bone marrow stromal cells. In normal circumstances, the interaction of all the cells mentioned, through stimuli such as growth factors, hormones, and cytokines, maintains the bone remodeling balanced. However, when there is an unbalance in bone remodeling, diseases can develop ranging from mild to severe. One example is osteoporosis, in which the bone resorbing process exceeds the bone formation process. When there is an unbalance in the bone formation process, this can result in sclerosing bone dysplasias namely osteopetrosis - Van Buchem disease and Camurati-Engelmann disease. Additionally, there are several metabolic bone diseases caused by defects in both bone resorption and formation processes. One of the most frequent from this last group is Paget's disease of bone (PDB), but there are also a number of rare conditions showing similarities to PDB:

- familial expansile osteolysis (FEO),
- expansile skeletal hyperphosphatasia (ESH),
- early-onset PDB (eoPDB),
- juvenile PDB (JPD),
- syndromal PDB condition named "inclusion body myopathy combined with PDB and frontotemporal dementia" (IBMPFD)[2].

Sir James Paget first described PDB (MIM 602080) in the 19th century as "osteitis deformans", a chronic inflammation of bone resulting in deformities. It is the second most common metabolic bone disease, after osteoporosis, with a prevalence of 2 to 5% in Caucasians over 55 years old. It affects more men than women, likely due to the larger mechanical loads on the bones of males[2–6]. PDB has a very unusual geographic distribution, presenting a different prevalence across different ethnicities, with the highest prevalence in European descent patients[2]. Within Europe, the highest prevalence belongs to United Kingdom, Spain, Italy, and France. It is also relatively common in people of European descent who have migrated to Australia, Canada, New Zealand, South Africa, and United States[7,8]. PDB occurs rarely in other parts of the world such as Africa, Middle East, and Asia[2]. PDB most likely originated in Britain and spread to other parts of the world as the result of migration and genetic admixture[7]. Furthermore, the incidence of PDB appears to be decreasing over the last 25 years, partly due to changes in the ethnic makeup of the population that result from the influx of migrants from low-prevalence regions such as the Indian subcontinent and the Far East.

Several PDB patients are diagnosed incidentally when being examined for other reasons since it is asymptomatic in approximately 80% of cases[2]. The characteristic focal bone lesions with accelerated bone turnover can be detected on bone scans using radionuclide-labeled bisphosphonates or on radiologic film with dual emission radiograph absorption, being the most sensitive method to detect pagetic lesions[2,8]. This method can be used to follow the activity of the disease in these patients[8]. Biochemical changes related to PDB include high levels of bone resorption markers (eg. urinary NTX [N-terminal telopeptide excretion]) and elevated levels of bone formation markers (eg. serum alkaline phosphatase), which are routinely used in clinical practice to make a diagnosis[2,9]. PDB diagnosis is performed based on the search for typical radiological features, measurement of serum alkaline phosphatase and a bone scintigraphy[10,11].

PDB develops in three consecutive phases. Initially (phase 1) there is an increased bone resorption (accelerated bone turnover) that gives an osteolytic appearance. Secondly (phase 2), mixed osteolytic and sclerotic features occur. Lastly (phase 3), sclerotic marks in the affected bone appear, which have a disorganized bone appearance (named "woven bone" or "cotton-like bone"). This affected bone has a reduced mechanical strength, placing patients at increased risk of developing bone deformities and pathologic fractures[2,7]. These lesions can occur in either just one bone

(mono-ostotic) or multiple bones (poly-ostotic) localized mainly in axial bones. In the majority of cases, the skin surface of the affected site present redness and warmth, since these lesions are highly vascular, and once a lesion is formed new lesions rarely develop. The most frequently affected bones are the pelvis, femur, (lumbar) spine, skull, and tibia. Additionally, and to a lesser extent, the knee, elbow, phalanges, and calcaneus can also be affected[2].

Complications can develop in PDB patients, including, fractures, bone pain, secondary osteoarthritis, deafness, spinal stenosis, nerve compression syndromes, and also heart failure due to the increased blood flow during active PDB (which leads to the patient death). Also, a subset of PDB patients develop osteosarcoma, suggesting that these patients may have an increased risk for the development of this malignancy when compared to the general population[2,12].

The genetic architecture of PDB is not yet fully understood for the majority of cases, however, it is consider to be a multifactorial complex disorder resulting from the combination of environmental and multiple genetic factors that, individually or epistatically, contribute the disease etiology[7].

1.2 Environmental factors

Several environmental factors have been suggested as possible triggers for PDB, including low dietary calcium intake during childhood, vitamin D deficiency, exposure to environmental toxins, repetitive mechanical loading of affected bones, a rural as opposed to an urban lifestyle, exposure to cattle, and chronic infection with measles, canine distemper or respiratory syncytial virus[2,5,7].

The most widely studied environmental exposure is paramyxoviral infection, although evidence of a viral etiology for PDB remains controversial[7]. Involvement of viral factors arose from the observation of paramyxoviral-like nucleocapsid particles (e.g. inclusion bodies) in the nucleus and cytosol of pagetic OCL[2]. Additionally, it has been suggested in 1960 that the introduction of immunization programs for measles and canine distemper virus is a possible reason for the reduction in PDB's incidence. However, this not only occurred too recently to account for a reduction in prevalence and severity of PDB patients born in the 1930s and 1940s (assuming that a slow viral infection picked up in childhood was indeed the cause of the disease), but also this would not be possible since the disease has a late age-at-onset[7].

In addition, there are reports of PDB cases that are due to arsenic acid (abundant in the water of a mill), which in turn resulted in a higher prevalence of PDB in inhabitants from Lancashire (Britain)[2,5].

Known genetic mutations associated to PDB appear to predispose individuals to this disease but require co-expression of an environmental factor in OCL precursors for the development of a robust "pagetic phenotype". In support of this hypothesis, Gennari *et al.* recently reported that there is a strong association between environmental factors (probably associated with a persistent animal contact) and the development of severe PDB in patients with a genetic predisposition to this disease (genes that increase the risk to familial PDB), again supporting PDB to be a complex disorder[13].

1.3 Genetic factors

In a subset of PDB patients, genetic risk factors play an important role in pathogenesis[14–16]. Familial clustering is frequently observed since 15 to 40% of patients have at least one affected first-degree relative[4,5,17,18]. Also, individuals with affected first-degree relatives have a sevenfold increased risk of developing PDB[14,19–21]. Since PDB is frequently asymptomatic prevalence of a familial aggregation could be underreported. However, it has been described that patients with a positive family history have a higher probability to manifest symptoms, deforming bone diseases and/or bone pain, when compared with patients with a negative family history[14,19,20]. Additionally, an earlier onset of PDB has been reported in family members when compared to sporadic cases[20]. One possible explanation for this is that relatives are more aware of PDB symptoms and seek medical attention at earlier ages[14]. Patients with a known family history will be called from now on as "familial PDB" and those with a negative family history as "sporadic PDB"[2].

Siris *et al.* reported a higher risk in siblings when the mother was affected versus an affected father, nonetheless a previous study could not find evidence for this[14,20]. One possible explanation for this higher risk is that there might be maternally transmitted factors, which could be either environmental factors (e.g. a virus), cytoplasmic factors, or imprinting of susceptibility genes[14].

In PDB families, segregation analysis showed that more than half of family relatives over the age of 55 had inherited the disease. Moreover, PDB has an equal incidence in males and females. Both these evidences point for PDB to be inherited in

an autosomal dominant fashion with highly variable penetrance, which is consistent with previous reports[8,15,21,22].

Several studies have shown that PDB might be caused by a combination of rare, high-penetrance variants in genes, such as *SQSTM1*, and other common variants in genes such as *CSF1*, *TNFRSF11A*, and *TM7SF4* (amongst others, see Table 1 and 2). These individually are not sufficient to cause the disease but can act together to increase the risk of developing it[7,23,24]. Also, the risk of developing PDB increases with an increased number of risk alleles carried[25].

Table 1. Summary of genes and *loci* identified using linkage analysis in PDB.

***Locus* name**	**Gene**	**Chromosome band**	**Type of original study**	**Associated SNP**	**Reference**
PDB1	***HLA***	6p21.3	Linkage	.	Tilyard *et al.*[26]
PDB2	***TNFRSF11A***	18q22.1	Linkage	rs3018362	Cody *et al.*[27]
PDB3	***SQSTM1***	5q35	Linkage	.	Laurin *et al.*[19]
PDB4	**.**	5q31	Linkage	.	Laurin *et al.*[19]
PDB5	**.**	2q36	Linkage	.	Hocking *et al.*[4]
PDB6	***OPTN***	10p13	Linkage	rs1561570	Hocking *et al.*[4]
PDB7	**.**	18q23	Linkage	.	Good *et al.*[28]
.	***VCP***	9p13.3	Linkage	rs565070	Kovach *et al.*[29]

Table 2. Summary of associated genes and *loci* in PDB.

Gene	**Chromosome band**	**Type of original study**	**Associated SNP**	**Reference**
CSF1	1p13	GWAS	rs484959	Albagha *et al.*[24]
CaSR	3q21.1	Candidate gene	.	Dónath *et al.*[30]
ESR1	6q24-q27	Candidate gene	.	Dónath *et al.*[30]
NUP205	7q33	GWAS	rs4294134	Albagha *et al.*[23]
TM7SF4	8q22	GWAS	rs2458413	Albagha *et al.*[23]
TNFRSF11B	8q24	Candidate gene	.	Wuyts *et al.*[31]
RIN3	14q32	GWAS	rs10498635	Albagha *et al.*[23]
PML	15q24	GWAS	rs5742615	Albagha *et al.*[23]
GOLGA6A	15q24	GWAS	.	Albagha *et al.*[23]

PDB is characterized by focal and disorganized increases in bone turnover, with highly localized areas of increased bone resorption (lytic phase) coupled with a high rate of bone formation with the primary cellular abnormality residing in OCL. This is accompanied by other abnormalities, such as marrow fibrosis and increased vascularity of bone[7,32,33]. OCLs, the primary affected cells in PDB, are increased in number and size, and express a "pagetic phenotype" that distinguishes them from normal OCLs[33]. "Pagetic" OCLs contain nuclear inclusion bodies, which are microcylindrical structures that have been linked to viral nucleocapsids. Although the identity of these inclusions has not been established until now, another hypothesis is that these inclusions might represent protein aggregates that are not degraded, similar to those observed in neurons from patients with neurodegenerative diseases. This is supported by the increasing evidence that PDB may be associated with a protein degradation system of autophagy dysregulation and the fact that OCL nuclear inclusions almost identical to those seen in PDB patients have been observed in mice carrying the *SQSTM1* (*sequestosome 1*) P394L mutation (equivalent to the human P392L mutation)[7].

Histologically, OCL are the primary affected cells in this disease. They are numerous, enlarged, hypermultinucleated, resistant to apoptosis and hyperactive, probably due to hypersensitivity for RANKL (receptor activator of nuclear factor-kB ligand), 1,25–$(OH)_2D_3$ (1,25–dihydroxyvitamin D3), and $TAF_{II}17$ (TATA-binding protein–associated factor [17 kDa], a vitamin D receptor binding protein)[2,34]. The $1,25(OH)_2D_3$ hyper-responsivity results from elevated levels of VDR (Vitamin D Receptor) coactivator and TAF12 (formerly TAF_{II}-17) in OCL[33]. Nuclear and cytoplasmatic inclusion bodies can be found in pagetic OCL[2]. OCL in PDB secrete high levels of IL-6 (Interleukin 6), which are detectable in marrow plasma and peripheral blood from PDB patients[33].

In sporadic PDB, mutations in *SQSTM1* cause a severe phenotype, with more affected bone and an earlier age of onset, than in patients without this mutation. However, similar phenotypes were observed in familial PDB with and without mutations in this gene. Furthermore, it was recently seen that PDB osteoclasts (with or without P392L mutation) from patients have the same phenotype including more nuclei per cell, apoptosis resistance and increased resorption activity. This raises the possibility that other factors (e.g. environmental and other genes) are involved in PDB susceptibility[2].

1.4 Polymorphisms

Genetic variation or polymorphism occurs where there is a variation in the deoxyribonucleic acid (DNA) sequence, being classified into several categories, including single-nucleotide polymorphisms (SNPs), insertions and deletions (InDels) and structural variants[35,36]. It is believed that genetic variation is the major factor that contributes to human diversity[36].

A SNP is a genetic variation that occurs in a single base of a DNA sequence, and where one of the four nucleotides is substituted for another. The minor allele frequency (MAF) must be greater than 5% in a given population to be considered a SNP (if the frequency is less than 1% it is called a rare variant). This is the most frequent type of variation in the human genome[36,37].

Insertions and deletion (InDels) are genetic variations where there is an insertion or deletion of at least one base from the reference DNA sequence. Several InDels have been detected over the past decade in the human genome, however their number is smaller than the number of SNPs and small InDels (ranging from 1 to 10 000 base pairs) cause similar phenotypic effects than SNPs[35]. Similar to SNPs, some InDels are localized in functionally important *loci*, likely to influence proteins and/or their regulation, and ultimately lead to human traits and diseases. However, InDels are more difficult to detect, validate and genotype, when compared to SNPs and so they are less studied[35].

1.5 Genetic variants that contribute to PDB risk

1.5.1 Linkage studies

Linkage is a genetic phenomenon described as the tendency of genes or other DNA segments at a specific genomic region to be inherited together on the same chromosome, as a consequence of their physical proximity[38].

Linkage analysis is designed to identify regions of the genome that contain genes that predispose to a trait/disease, testing for cosegregation between a well-characterized polymorphic genetic marker and an unknown *locus* influencing the disease susceptibility, using Affected Sibling Pairs (ASP) or extended families[38,39]. Genetic linkage methods can be parametric (model-based) and non-parametric (model

free). The former is used when the genetic model of the disease is specified, being more often used in Mendelian diseases, whereas the latter is used when there is no clear mode of inheritance of the disease (such as with many complex disorders), or when I do not know the allele frequencies and/or penetrances are not known[38].

The logarithm of the odds (LOD) score, proposed by Morton in 1955, is a function of the recombination fraction or chromosomal position measured in centimorgans (cM), being a useful measure of linkage. High LOD scores constitute evidence for linkage and low scores represent evidence against. It was suggested that linkage is excluded when the LOD is less than -2, and a LOD score of 3 or higher indicate a significant evidence for linkage[38]. However, in complex diseases, like PDB, where several genes could contribute to increase the risk of the disease, the stringency of the criteria should be different, and a modest maximum LOD score is expected and should not be ignored[38]. The limited success of linkage analysis for complex diseases is in part due to studies being too small to detect genes of modest effect. The sample size necessary to detect linkage to genes with a genotype relative risk of less than 2 could be unachievable[38].

Linkage analysis can only identify large regions and even if there is a strong causal gene within the linkage peak, such regions frequently contain hundreds of genes, many of them biologically reasonable candidates in a complex disease. A good approach to narrow the region of interest is to perform association analysis to fine-map the linkage peak[38].

Suggestive evidence for linkage in PDB was first reported by Tilyard *et al.* for *HLA* at 6p21.3 (also named *PDB1 locus*) – Table 1[26]. However, no other study has confirmed the *HLA* linkage with PDB, indicating that the linkage signal may have been a false positive or the gene may be of minor importance in the disease etiology[14,19,40,41].

Based on a study for FEO (a bone disorder similar to PDB previously mentioned), where evidence of linkage has been shown at 18q21.2-21.3, Cody *et al.* performed a linkage study in this region in two PDB families in which the authors found evidence for linkage with that same *locus*[27,42]. *TNFRSF11A* (*PDB2 locus*) maps within this region, and encodes a receptor activator of nuclear factor κB (RANK) that regulates OCL activity[16]. However, subsequent reports are contradictory. Some studies report a strong linkage signal at 18q[27,43] while other studies found no evidence of linkage at this *locus*[19,22,4,19]. In addition, a whole-genome linkage study performed in a large pedigree

did not show any evidence for linkage at *PDB2*, identifying a novel susceptibility *locus* at 18q23 (*PDB7*) near of *TNFRSF11A*[28].

Laurin *et al.* conducted a whole genome linkage study in 24 large French-Canadian families with PDB, detecting evidence for linkage at both 5q35 (*PDB3*) and 5q31 (*PDB4*)[19]. The former contains *SQSTM1*, which is the only disease-causing gene consistently reported for PDB. Mutations in this gene are reported in 40-50% of familial PDB and 2.5-10% of sporadic PDB, usually following an autosomal dominant mode of inheritance[2,7,17].

SQSTM1 is a multifunctional protein with 440 amino acids, distributed by nine protein-interacting domains. The main functions of SQSTM1 are proteasomal degradation of proteins, acting as scaffold protein in the RANKL, IL1, nerve growth factor and TNFα-induced NFκB signaling pathways, autophagy, and apoptosis. Moreover, SQSTM1 is an important protein in the RANKL/RANK/OPG-NFκB axis. At least 28 mutations in *SQSTM1* have been reported in about one third of PDB families as well as in about 9% of sporadic PDB[2,25]. All mutations identified in *SQSTM1* are located within and around the ubiquitin-associated (UBA) domain of the protein. The most frequent mutation found in familial and sporadic PDB is $p62^{P392L}$, a mutation causing an amino acid substitution from a proline (P) to a leucine (L) at position 392 of the coding sequence[2,33,44]. This mutation induces the activation of human osteoclasts[44].

Additional linkage studies performed by Hocking *et al.* demonstrate evidence for linkage at 2q36 (*PDB5*), 10p13 (*PDB6*), and also 5q35 [4]. *OPTN* (optineurin), a homolog of NFκB (nuclear factor kappa B) essential modulator (NEMO) implicated in the NFκB signaling pathway and autophagy, is located in *PDB6*[2,7].

In families reported to have linkage in *PDB2* and *PDB7* regions, most of the patients also carried *SQSTM1* mutations[40,41]. Furthermore, when Lucas *et al.* reanalyzed linkage data excluding patients with *SQSTM1*, mutations there was only evidence for linkage at 10p13 (*PDB6*), while at 2q36 (*PDB5*) the linkage signal disappeared almost completely. There are two possible explanations for this: first, *PDB2*, *PDB5* and *PDB7 loci* contain modifier genes that interact with *SQSTM1* to cause the disease or secondly, the former results are false positives[40,41].

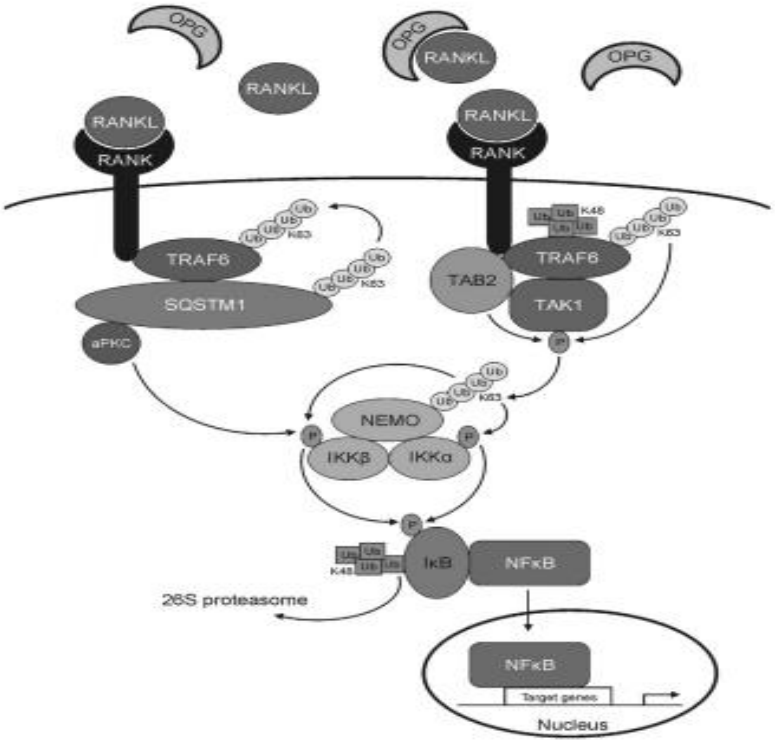


Figure 1. RANKL/RANK NFκB signaling pathway.

Another linkage study performed by Kovach *et al.* in a family with PDB showed linkage at 9p13.3. Later, a genome-wide association study (GWAS) confirmed the association between *VCP* (valosin-containing protein), localized in this region, and PDB[12,29]. This gene expresses a protein that has been linked to the autophagy mechanism and may be involved in cellular and structural functions of muscle and/or bone cells[12,29].

To sum up, all these findings reinforce the notion that PDB is genetically heterogeneous and multifactorial[19,27].

1.5.2 Association studies

From 1970 to 1990, linkage analysis was the dominant approach in the investigation of genetic risk variants in families, however association studies are more powerful in the detection of genes with a modest/small effect[37,45].

Association studies allow the identification of susceptibility genes, detecting common variants with weak/moderate effect on the phenotype at the population level[45]. These common genetic variants, in particular SNPs, usually have a MAF above 5%[46].

Evidence for association between a marker and PDB exists if there is linkage disequilibrium (LD) between that marker and the causative functional variant or if the marker itself corresponds to the causal gene variant[47]. LD is a genetic phenomenon described as two or more genetic variants that are inherited together on the same haplotype more often than expected by chance alone in the population under study[38].

There are many sampling approaches to perform an association study, being the most common either a case-control or a family-based format. In the case-control approach, an unrelated group of patients is compared to a group of matched controls. In the family-based studies patients and their family members are collected and compared[47–49]. Family-based studies are more robust because they are less affected by population stratification, whereas case-control studies are more powerful, easier to collect and genotype[48].

Association studies can be performed on a candidate gene basis or as a GWAS[7,12]. In the first approach, genes are selected based on their function (e.g. role in bone metabolism or a relationship to bone diseases for PDB). The genome-wide approach uses variants across the entire genome independently of gene function, increasing the probability of discovering novel unbiased susceptibility variants[12].

GWAS carried out for PDB identified six potential susceptibility *loci* (Table 2): 1p13 (*CSF1*)[24], 7q33 (*NUP205*)[23], 8q22 (*TM7SF4*)[23], 9p13.3 (*VCP*)[12], 14q32 (*RIN3*)[23], and 15q24 (this region contain two genes, *PML* and *GOLGA6A*)[23].

The function of these genes appears to be related with PDB. *CSF1* encodes for the macrophage-colony stimulating factor (M-CSF), which plays a key role in osteoclast formation and survival[2,7,24]. *NUP205* encodes the nucleoporin 205 kDa protein, which has a role in the nucleus to cytosol transport of proteins such as ALFY (Autophagy-Linked FYVE domain-containing protein/WD repeat and FYVE domain containing 3), which interacts with SQSTM1[2,23]. *TM7SF4* (transmembrane 7 super-family member 4) encodes the dendritic cell-specific transmembrane protein (DC-STAMP), which functions in the fusion of OCL precursors to form mature OCL[2,23]. *RIN3* regulates the vesicular trafficking via interaction with small GTPases, and could have a function in the bone resorption since small GTPases have an important role in vesicular trafficking and in OCL function[2,23]. *PML* (promyelocytic leukemia gene) plays an important role in

the TGF-β (tumor growth factor-β) signaling pathway, which regulates bone remodeling suggesting that *PML* may have an influence on the coupling between bone resorption and bone formation through TGF-β[2,23]. *GOLGA6A* (golgin A6 family member A) is located on the same associated regions as *PML* and its function in bone metabolism is not yet known, but other mutations in other golgin proteins have been associated with skeletal dysplasia and a severe form of osteoporosis. GOLGA6A protein has a known function in the Golgi apparatus and in membrane fusion[2,23].

There are also other association studies based on candidate genes, which report an association of *CaSR*[30], *ESR1*[30] and *TNFRSF11B*[31,50] in genetic susceptibility for PDB.

CaSR has an important role in the regulation of extracellular calcium that could be important for the changes that occur in the bones of PDB patients[30]. *TNFRSF11B* encodes osteoprotegerin (OPG), a key regulator of osteoclastic bone resorption. OPG is a member of the TNF (tumor necrosis factor) receptor superfamily that lacks an intracellular domain and acts as a decoy receptor for RANKL, thereby inhibiting osteoclast differentiation and bone resorption[7]. In addition, *TNFRSF11B* seems to be sex-related, being more associated to women in a Belgian population study[51]. In the Donáth *et al.* study there is also a sex specific effect (women more affected than men) possibly due to polymorphisms in *ESR1* (estrogen receptor 1)[30].

Several reports suggest that abnormalities of OCL formation in PDB could be due, in part, to the cytokine IL-6. Elevated levels of IL-6 were detected in plasma and in conditioned media from pagetic bone marrow cultures when compared with control cultures[2,12,40,52]. Moreover, the vitamin D receptor (*VDR*) gene has a function in calcium and bone metabolism, since vitamin D is needed for a normal bone mineralization, absorption of calcium from the gut, control of calcium and phosphate homeostasis and regulation of parathyroid hormone secretion[30]. Finally, *TNFSF11* encodes RANKL, having an important role in NFκB-induced osteoclastogenesis[12]. However, no SNPs in *IL6*[12], *VDR*[30] or *TNFSF11*[12] were found to be associated with PDB[2,12,40].

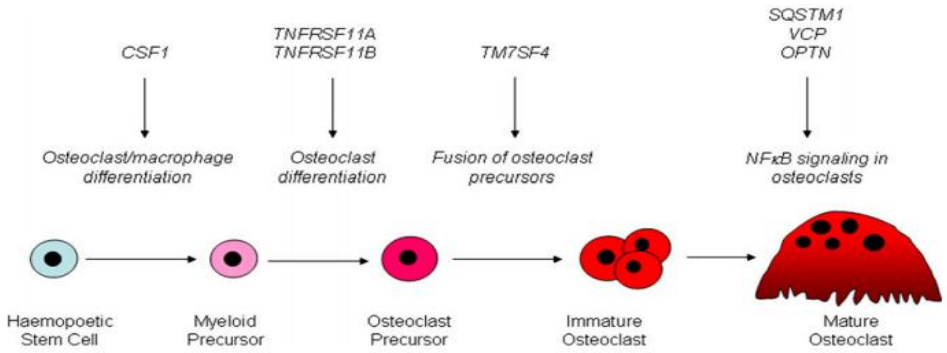


Figure 2. Genes that predispose to PDB play key roles in osteoclast differentiation and function. OCL differentiation has multiple steps with several genes involved in its pathway. *CSF1* is essential for OCL and macrophage differentiation. *TNFRSF11A* and *TNFRSF11B* encode RANK and OPG, respectively, playing both an important role in OCL differentiation and function. *TM7SF4* is needed for OCL precursors fusion; *SQSTM1*, *VCP* and *OPTN* have an important function in the regulation of NFκB signaling and autophagy[7].

Linkage and association data altogether indicate that *CSF1*, *OPTN*, *TNFRSF11A*, *TM7SF4*, *NUP205*, *RIN3*, *PML* and *GOLGA6A* are putative players in PDB genetic risk[2,17]. However, more studies are warranted to identify other causal variants that explain the missing heritability in PDB risk.

1.6 Next-generation sequencing (NGS)

GWAS helped to identify more than 2,000 common variants that play a role in complex disease susceptibility and provided many new clues about disease biology. However, common variants tipically explain a small proportion of the genetic variance and much of the genetic contribution remains unexplained. Partially, this might be due to rare variants (MAF < 5%) that are not detected with GWAS but that could play an important role in human diseases[25,46]. These rare variants may be independent of each other and confer a detectable risk of developing a disease[25]. Many diseases are caused by a combination of highly penetrant rare variants and common variants. Reports show that rare variants have an important role in complex diseases[53]. NGS enables the detection of these rare variants, complementing the investigation in the complex diseases field[25,46,53].

NGS is a high-throughput parallel-sequencing approach that produces hundreds of thousands/millions of short-reads (approximately 100-500 bp) in a short time[46].

These short reads are then aligned against a reference genome to identify where sequenced individuals vary[46].

There are two main types of NGS: Whole-genome sequencing (WGS) and whole-exome sequencing (WES). The first has a greater cost and an inherent complexity in analysis. WES is a robust strategy for discovering novel rare variants associated to complex diseases since the exome (1-2% of the genome, which encodes for proteins) represents a highly enriched subset of the genome, in which to search for variants with large effect sizes[46,54–56].

A large fraction of rare, protein-altering variants, such as non-synonymous (NSV) or stop-gain/loss single-base substitutions or small InDels, are predicted to have functional consequences and/or to be deleterious[56]. First successful cases of NGS carried out in rare monogenic disorders were Miller syndrome[57], Freeman-Sheldon syndrome[58] and Schinzel-Giedion syndrome[59]. There is an increasing number of studies using WES to identify genes associated with complex diseases[46,60]. Although WES is a robust approach to search for disease-related variants, it has a high error rate[61].

WES analysis has several technical constrains when identifying risk alleles, depending on their mode of inheritance. Genes that cause recessive and *de novo* dominant diseases are easier to identify due to low number of genes shared between affected individuals, decreasing the number of possible candidate genes[55]. Identification of genes that cause dominant diseases is more challenging due to several reasons such as a small number of family members and consequently a higher number of candidate heterozygous variants, or the absence of disease-causing variants in the mapped region[55]. In other cases, no variants are identified since the causal mutation resides outside the exome.

So far, polymorphisms in a total of 13 genes (Table 1 and 2) have been associated to increased PDB risk. Since *SQSTM1* mutations can only explain PDB heritability in a small subset of patients, the search for novel rare causal variants is now warranted. Unraveling the genetic background of PDB will be essential to understand its pathogenesis, and it might make early detection and treatment for individuals at risk a reality[2].

2. Objectives

Our goal is to identify the gene(s) that cause PDB in an extended Portuguese multiplex family from *Alentejo*.

More specifically, the project was divided in the following tasks:

1. Whole-exome sequencing (WES) of six family members;
2. Bioinformatic analysis of WES data based on the Genome Analysis Toolkit (GATK) Best Practices workflow from Broad Institute, for human exome sequencing;
3. Selection of candidate variants - present in the affected relatives and absent in the unaffected relatives – as well as analysis of their potential functional impact using *in silico* methods (e.g. SIFT, PolyPhen);
4. Technically validate the candidate variants by Sanger sequencing and test their segregation in all available family members;
5. Test the presence and segregation of these mutations in a second PDB multiplex family also from *Alentejo*.

3. Methods

3.1 DNA collection and extraction

7.5 mL of whole blood was collected using sodium Ethylenediaminetetraacetic Acid (EDTA) tubes. For each individual, the medical team collected two blood EDTA tubes, one for DNA extraction and another for pellet isolation. This was the preferred DNA source for the study performed since it was the source already available, and also it allows the extraction of large amounts of DNA required to perform genomic assays[62].

DNA extraction was carried out using Genomic DNA Extraction kit (RBC Bioscience Corp., New Taipei City, Taiwan) that is designed to purify genomic DNA from 50 μL to 10 mL of blood.

When I started my master project, family 1 DNA samples (except for individuals III.4-080001 and IV.3-090095) were already extracted. The kit used was Nucleo Spin Blood XL kit (Macherey-Nagel, Germany) according to the manufacturer's protocol. I only collected and extracted the blood samples from family 1 III.4-080001 and IV.3-090095 and family 2 individuals.

In summary, the protocol consisted in:

1. Pipet up to 400 μL of whole blood to a 2.0 mL microcentrifuge tube;
2. Add three times the sample volume of RBC lysis buffer (to remove red blood cells) and mix by inversion 10-15 times (no vortex);
3. Incubate for 10 minutes at room temperature;
4. Centrifuge at 3.000 g for 7 minutes;
5. Remove the supernatant, but retain about 50 μL of residual buffer to resuspend the white cell pellet by vortexing;
6. Add up to 400 μL cell lysis buffer (to lyse cell membrane) to the tube and mix by vortexing;
7. Incubate at 60ºC for 20 minutes, until the sample lysate was more or less clear. During incubation invert the tube every 4/5 minutes;
8. Add 125 μL of protein remove buffer (to digest proteins) to the sample lysate and mix immediately by vortexing for 10 seconds;

9. Incubate on ice for 10 minutes;
10. Centrifuge at 13.000 rpm for 5 minutes;
11. Transfer the supernatant from Step 11 to a 1.5 mL microcentrifuge tube;
12. Add 400 μL isopropanol (for DNA precipitation) and mix well by inverting;
13. Centrifuge again at 13.000 rpm for 5 minutes;
14. Discard the supernatant and add 400 μL of 70% ethanol (to improve DNA precipitation) to wash the pellet;
15. Centrifuge at 13.000 rpm for 5 minutes;
16. Discard the supernatant and air-dry the pellet for 20 minutes;
17. Add 50-100 μL of water and incubate at 60ºC for 30-60 minutes to dissolve the DNA pellet. During incubation, tap the bottom of tube to promote DNA rehydration.

For individual IV.9-090044, DNA extraction was made from the pellet since the whole blood sample was no longer available. This required two additional steps before the DNA extraction. First, the blood sample was centrifuged to split the white blood cells from the rest of the cells. Then, it was added PBS (Phosphate buffered saline, to ressuspend the pellet) up to 7.5 mL.

3.2 DNA quantification

DNA concentrations were quantified using the NanoDrop 2000. Briefly, 1 μL of each sample was pipetted onto an optical pedestal (receiving fiber). Another fiber is then brought into contact with the first one by closing the NanoDrop's arm, forcing the 1 μL solution to fill the gap between the two fiber optic ends in order to be measured. A source light is then directed to the sample, and the spectrometer analyses the light after passing through the solution based on the amount of radiation absorbed. The data is then stored using a computer NanoDrop 2000/2000c software. Before measuring the samples, 1 μL of MilliQ water is used as a blank measurement. This spectrophotometer uses a spectra range from 190 to 840 nm, giving accurate estimates of the template concentration in ng/μL. DNA has optimal absorption at 260 nm, due to the structure of the aromatic rings of their bases[63]. Thus, DNA concentration can be estimated based on the amount of absorbed radiation at 260 nm[62,64]. The Lambert-Beer equation is applied to correlate the absorbance with concentration

$$A = \varepsilon.l.C$$

where A is the absorbance, ε is the wavelength-dependent molar absorptivity coefficient ($ng^{-1}cm^{-1}\mu L$), l is the path length (cm), and C is the nucleic acid concentration ($ng\mu L^{-1}$) [64]. Standard values for the molar absorptivity coefficient of double- and single-stranded DNA (0.020 $ng^{-1}cm^{-1}\mu L$ and 0.027 $ng^{-1}cm^{-1}\mu L$ respectively, when exposed to Ultra Violet light (UV) at 260nm) were applied for the concentration estimation[64].

The purity of the sample can then be roughly determined using the A260/A280 ratio. The DNA is considered pure if this ratio is 1.8[62,63]. For RNA, it is considered pure when the A260/A280 ratio is 2.0[62,63]. If this ratio is significantly different from 1.8, then the DNA sample is contaminated, either by RNA (if the ratio is closer to 2.0) or by proteins or other contaminants (if it is below 1.8), since the latter will absorb radiation at 280 nm[62,63].

For the WES and validation assays, all DNA samples were diluted to a final concentration of 40 ng/uL (minimum concentration required for Beijing Genomics Institute – BGI and 10 ng/uL, respectively (working solutions).

3.3 Electrophoresis

Gel electrophoresis was mainly used to examine DNA quality, fragment size and PCR product specificity. The DNA fragments always migrate from the cathode (negative pole) to the anode (positive pole), because DNA carries a negative charge due to its phosphodiester backbone. As the DNA migration in the gel matrix occurs, the fragments are separated according to size - small fragments migrate faster and run further in the gel compared to larger ones. For optimal resolution the agarose concentration varied from 1-3% depending on the size of the PCR products, allowing for maximum fragment separation. The gels were made by dissolving agarose powder (NZYTech, Lisboa, Portugal) in 1x Tris-Acetate-EDTA (TAE) (Sigma, Missouri, United States of America), followed by heating the mix in a microwave until it becomes transparent (meaning that the agarose powder is fully dissolved). GreenSafe Premium (NZYTech, Lisboa, Portugal) was added at a final volume of 6 μL in 100 mL of agarose gel. The whole solution was poured into a sealed plate and allowed to cool and set. Once the gel was set, the gel was immersed in a 1x TAE buffer solution. Loading buffer (3.5 μL) was added to each sample (6 μL) and these were loaded into the wells with a molecular weight ladder running in parallel (6 μL of 1x NZYDNA Ladder VI

[NZYTech, Lisboa, Portugal]). This ladder provided an approximate quantification of the fragment's size. Moreover, we used a negative control (using all PCR reagents but instead of DNA we added water) to guarantee that there was no contamination. The voltage was set at a constant value (usually at 100 V) for the length of time required (which can be visually checked by the migration of the loading buffer). As the products migrate the GreenSafe intercalates with the DNA, and consequently allowing its visualization using an UV transilluminator (GenoSmart gel documentation system, VWR International) and photographed.

3.4 Bioanalyzer

The bioanalyzer system is an electrophoretic assay based on traditional gel electrophoresis principles that have been transferred to a chip format. This platform provides sizing, quantitation and quality control of DNA, dramatically reducing separation time as well as sample and reagent consumption. 12 samples (three PDB patients and nine positive controls since the PDB DNA samples were older) were run using bioanalyzer. This analysis was carried out in this platform to make sure that the DNA had high quality for the WES assay, which is more sensitive than Sanger sequencing.

The total amount of DNA must be between 0.5-50 ng/μL for an accurate determination of DNA concentration with the Bioanalyzer. The chip accommodates sample wells, gel wells and one well for an external standard (ladder). During chip preparation, the micro-channels are filled with a sieving polymer and fluorescence dye, once the wells and channels are filled, the chip becomes an integrated electrical circuit. The 16-pin electrodes of the cartridge are arranged so that they fit into the wells of the chip. Each electrode is connected to an independent power supply that provides maximum control and flexibility, such that DNA is electrophoretically driven by a voltage gradient. The molecules are separated by size (smaller fragments migrate faster than larger ones) due a constant mass-to-charge ratio and the presence of a sieving polymer matrix. Dye molecules intercalate into DNA strands and the complexes formed are detected by laser-induced fluorescence. Data is translated into gel-like images (bands) and electropherograms (peaks). A standard curve of migration time versus fragments size is plotted with the help of a ladder that contains components of known sizes, and size is calculated from the migration times measured for each fragment in the

sample. Two markers are run with each of the samples bracketing the overall sizing range. The "lower" and "upper" markers are internal standards used to align the ladder data with data from the sample wells, to compensate for drift effects that may occur during the course of a chip run[65].

3.5 Whole Exome Sequencing (WES)

The exome of six family members (from family 1) with multiple PDB patients was sequenced using WES outsourced at BGI (Hong Kong, http://www.genomics.cn/en/index). This technique is highly complex, and involves multiple steps that may be divided into (1) genomic enrichment (selection of all exons with an additional 200 bp up and downstream of each target region), (2) sequencing (including library preparation), and (3) bioinformatics analysis[66].

The genomic enrichment step (step 1) is a multi-step procedure (Figure 3). Briefly, the qualified DNA sample is randomly sonicated into fragments with a base pair peak of 150 to 200 bp (in our case the fragments had 100 bp approximately). Afterwards, adapters are ligated to both ends of the resulting fragments. The adapter-ligated templates are purified by the AgencourtAMPure SPRI beads and fragments with insert sizes of about 200 bp are excised. Extracted DNA is amplified by ligation-mediated polymerase chain reaction (LM-PCR), purified and hybridized using the SureSelectBiotinylated RNA library (BAITS) for enrichment of the exonic regions.

Hybridized fragments are bound to the strepavidin beads whereas non-hybridized fragments are washed out after 24h. Captured LM-PCR products are subjected to Agilent 2100 Bioanalyzer to estimate the magnitude of enrichment. Because the aim of this analysis is to sequence the protein-coding part of the genome (exome), such targeted sequencing includes the enrichment of the target sequences[54]. Each captured library is then loaded on Hiseq2000 (Illumina's Solexa) platform.

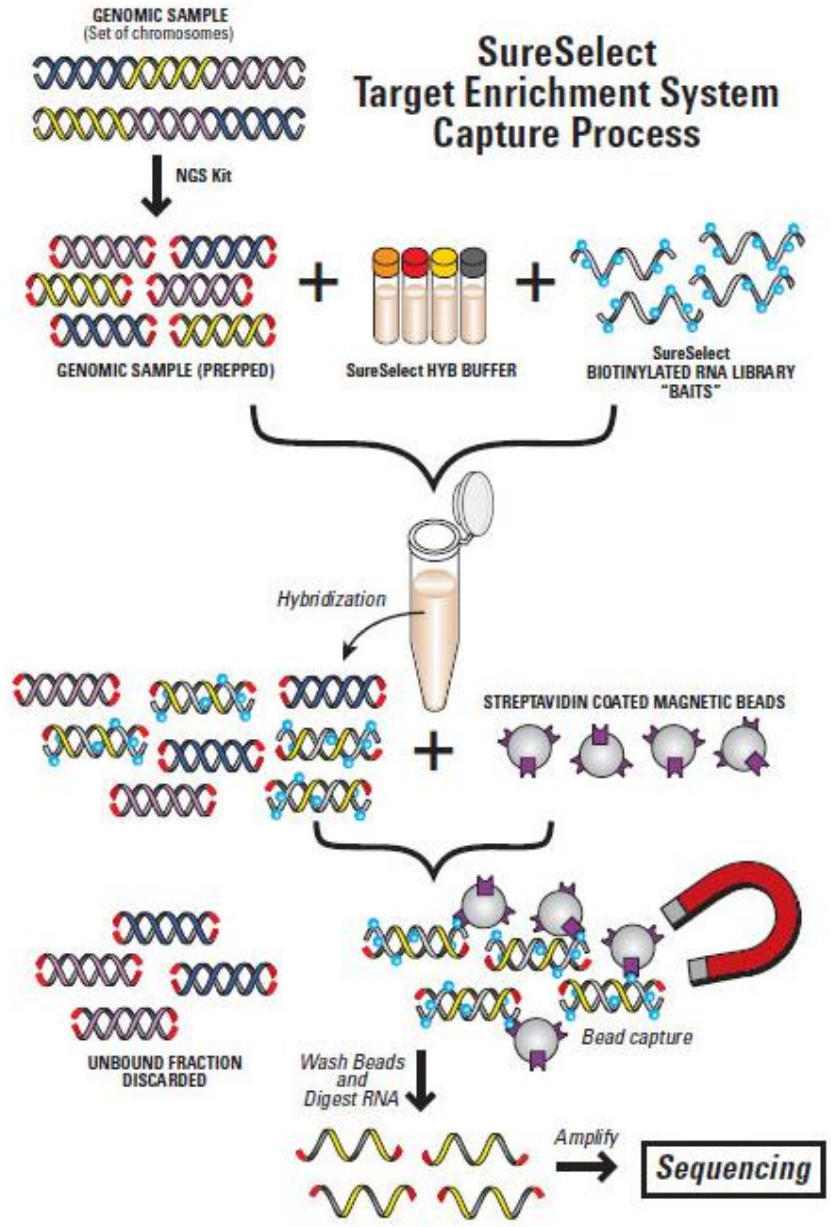


Figure 3. Steps of the exome capture used to sequence PDB family 1 (http://ncifrederick.cancer.gov/atp/genetics-and-genomics/laboratory-of-molecular-technology/lmt-protocols-and-resources/sureselect/background/).

Afterwards, high-throughput sequencing (step 2) was performed with an average fold-coverage of 60x approximately (recommended coverage value). Raw image files are processed by Illumina base calling software 1.7, using base calling default parameters that vary between 2 and 41 (these are standard values used by BGI). This base quality value (denoted as Q) allows for the calculation of the sequencing error rate (denoted as E, $sQ = -10\log_{10} E$, Table 3).

Table 3. Relationship between the sequencing error rate (*E*) and the quality value (*Q*). Three examples of sequencing error rates and the corresponding sequencing quality values (calculated using the formula above) are shown.

Sequencing error rate (*E*, %)	Sequencing quality value (*Q*)
5	13
1	20
0.01	30

The sequences of each individual are then generated with 90 bp paired-end reads. Illumina´s technology is based on clonally amplified templates coupled with cyclic reversible termination method with four fluorescent colors. First, one fluorescently modified nucleotide complementary to the template sequence is incorporated. After washing and imaging for detection of the incorporated nucleotide, a cleavage step removes the fluorescent dye and a novel incorporation step is performed. These steps are done in a cyclic manner, 72 or 100 times. It is possible to sequence from both extremities of the DNA template (paired-end sequencing[a])[54].

The final steps (step 3, bioinformatics analysis - see Figure 4) are the genome alignment, variant calling and data analysis.

The Hiseq 2000 is one of the most robust platforms with lower false positive rates, detecting the signal produced by the incorporation of nucleotides. Thus, for sequencing platforms using single molecule templates, the amount of starting DNA is

[a] Paired-end sequencing allows sequencing both ends of a strand by the ligation of adapters, containing attachment sequences, and sequencing primer sites (forward and reverse) into exonic DNA. Single-end sequencing on the other hand only allows sequencing of one strand because it only uses one sequencing primer (forward or reverse). In our study we used paired-end reads always to improve the confidence of our results.

lower and there is no PCR amplification step that could create artificial mutations and AT or GC-rich amplification bias. Templates, primers and polymerase enzymes are immobilized on a solid support before the sequencing reaction[54].

WES is already a useful alternative or complementary technique for molecular diagnosis. Its routine use leads to a rapid screening and fast identification of mutations in rare genetic disorders through sequencing of the coding region[54].

3.5.1 Bioinformatics analysis

We followed the Genome Analysis Toolkit (GATK) Best Practices workflow (Figure 4), from the Broad Institute, optimized for our human data (all the scripts used are in Appendix A). However, some tools were added to the pipeline to improve the analysis (such as Freebayes and Samtools mpileup used for the variant calling).

WES data enters a bioinformatics pipeline that includes data pre-processing (e.g. sequence aligned using BWA), variant discovery, and ultimately a variant annotation (e.g. *in silico* evaluation of variant function) (Figure 4). All data has been stored redundantly at IMM. We adhered to the Minimum Information about a Genome Sequence (MIGS) specification (http://gensc.org).

3.5.1.1 Alignment

The bioinformatics analysis begins with the sequencing data (raw data), which was generated from the Illumina pipeline. BGI eliminated reads that contained the sequence of the adapter, the low-quality reads (which have long stretches of "N"s) and reads in which unknown bases are more than 10%. This step produced the "clean data", originating "clean" FASTQ files used in the bioinformatics analysis. The generated reads (FASTQ files) were aligned to the human reference genome hg19/GRCh37 (available at http://hgdownload.cse.ucsc.edu/goldenPath/hg19/bigZips/) using the Burrows-Wheeler Aligner tool (BWA – version 0.6.1) leading to binary alignment files (BAM) (http://bio-bwa.sourceforge.net/). This tool identifies the best position where the reads match and then does a linear scan through all the potential hits for both paired-end reads.

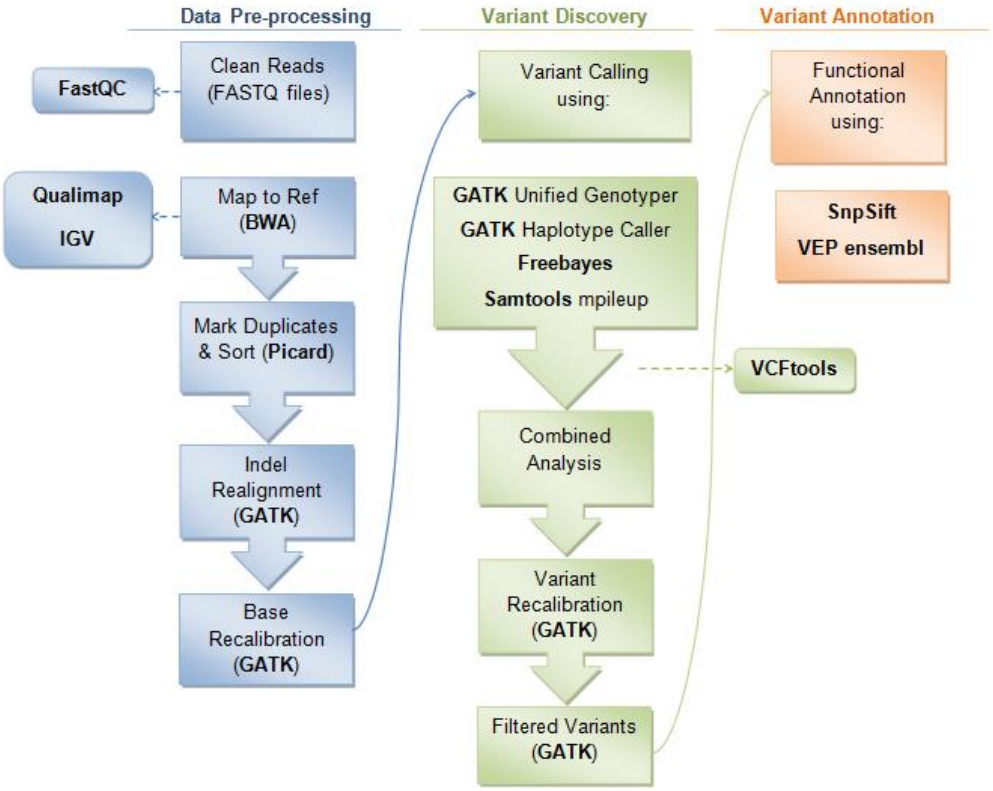


Figure 4. Bioinformatics analysis workflow. This pipeline is divided in three main sections: Data pre-processing, variant discovery and variant annotation. The first section is the alignment and base recalibration, starting with the FASTQ files and ending with the BAM files ready for further analysis. The second section is the variant calling, originating the VCF files which are used in the annotation phase. In the third section, we make the variant annotation using two different tools (SnpSift and VEP ensembl). Quality control of the FASTQ files and BAM files was made with FastQC and Qualimap, respectively, which were added steps to improve the confidence of our results.

BWA estimates the insert size distribution of both paired-end reads mapped and pairs them. After that, it aligns the unmapped reads whose mate pairs are already respectively aligned with higher confidence[67]. For each mapped read, BWA generates a mapping quality score (MAPQ), which varies between 0 and 60[67]. The higher the MAPQ score is, the smaller the probability of the alignment being incorrect. Finally, BWA records the results for each read as a CIGAR string (Figure 5)[67]. The BAM format files are needed to do other processes, such as fixing mate information of the alignment, adding read group information, marking duplicate reads and InDel realignment.

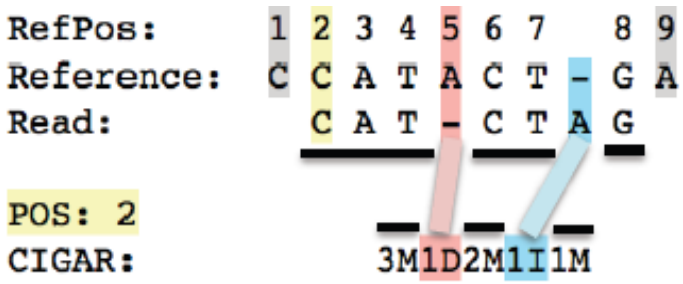


Figure 5. CIGAR string example. The POS indicates that the read aligns starting at position 2 in the reference. The CIGAR string indicates that the first three bases in the read sequence align with the reference. The next base in the reference does not exist in the read (one deletion). Then two bases align with the reference. The next base in the read does not exist in the reference (one insertion) and then one more base aligns with the reference.

3.5.1.2 Quality control (QC)

The sequencing reads were aligned with the reference genome sequence using BWA (explained above). Also, the duplicated reads (redundant information produced by PCR) were removed using Picard (http://picard.sourceforge.net/command-line-overview.shtml - version 1.111). This tool sort the reads present in the dataset according to chromosomal position and marks as duplicates all the CIGAR strings that match for alignments in the same position, except for the best read that has the highest sum of base qualities ($Q \geq 15$). Also, for an internal quality control, FastQC (version 0.10.1) was used to analyze the FASTQ files and Qualimap (version 0.7) to analyze the BAM files generated by BWA.

After these quality control steps, the BAM files were re-aligned with the GATK IndelRealigner. This step is crucial because InDels could lead to errors in the alignment itself originating mismatches. Local realignment around InDels reduces the number of mismatching bases.

Lastly, a base recalibration was performed using the GATK base quality recalibration tool. This identifies the number of mismatches in the sequence and uses the known variants (present in Mills and 1000 Genomes Project (GP) gold standard – a stringently curated list of InDels) to only take into account the novel genetic variation, and so discard most of the known genetic variation, identifying the probability of error for each base with a higher confidence for variants in databases. BAM files are at this

point ready for further analyses. A list of SNPs and InDels was derived from the BAM files based on unique matches with the reference genome. Two lists were created for analysis, one for SNP and another for InDels variants.

The quality of the genetic variants was also inspected visually using IGV (Integrative Genomics Viewer - http://www.broadinstitute.org/igv/ - version 2.3.32). IGV is a software that allows the visualization and interactive exploration of genomic data, including NGS data[68]. This tool permits the visualization of data in multiple genomic regions simultaneously in adjacent panels, "a scale of genome resolution from whole genome to base pairs"[68]. It also provides several reference genomes for different species, and provides the option of importing other genome references (but only FASTQ files containing chromosome or contig sequences)[68]. IGV supports several read alignment file formats, including SAM and BAM files.

3.5.1.3 Variant calling – SNPs and InDels

For the variant calling four of the most popular tools from academic and industrial settings were used[69]:

1) GATK Unified Genotyper (http://www.broadinstitute.org/gatk/gatkdocs/ org_broadinstitute_sting_gatk_walkers_genotyper_UnifiedGenotyper.html - version 2.4.3 and 3.1.1);

2) GATK Haplotype Caller (http://www.broadinstitute.org/gatk/ gatkdocs/ org_broadinstitute_sting_gatk_walkers_haplotypecaller_HaplotypeCaller.html - version 2.4.3 and 3.1.1);

3) Freebayes (https://wiki.gacrc.uga.edu/wiki/Freebayes - version 0.9.14);

4) Samtools mpileup (http://samtools.sourceforge.net/mpileup.shtml - version 0.1.19).

The differences between the two GATK tools are that Unified Genotyper calls separately SNPs and InDels by considering each variant *locus* independently, whereas in Haplotype Caller SNPs and InDels calls are performed with a local *de novo* assembly. Only bases that fulfill the criteria specified in the program are included, namely a minimum base quality value required to consider a base for calling and a good confidence call[70]. To determine the most likely combination of genotypes at each position in the genome, Freebayes uses short-read alignments for any individual and a reference genome (in our case GRCh37/hg19)[71]. Finally, Samtools mpileup needs the

bcftools to make the variant call because it is only capable of collecting the information of BAM files, computing the likelihoods of data given each possible genotype and stores the scores in the binary format (.bcf file). Then, Bcftools (from Samtools) makes the variant calling applying the prior likelihoods obtained by Samtools and doing the actual calling[72].

Briefly, all four tools used (GATK Unified Genotyper, GATK Haplotype Caller, Freebayes and Samtools) use a Bayesian formulation for picking the base that maximizes the posterior probability with the highest Phred quality score (this is the *Q* parameter previously described, which characterize the quality of each base call)[71,73,74]. In addition, for each individual, a combined analysis was performed, where the variants that are not common between the four variant calling tools were filtered out.

Transitions (Ti - A – G, C – T) are twice as frequent as transversions (Tv - A – C, G – T, A – T, C – G), so false positive SNPs should have Ti/Tv (Transition/Tranversion) around 0.5 for randomly assigned variations, such results could be the consequence of systematic sequencing errors, alignment artifacts and data processing failures[75]. For evaluating the specificity of novel SNP calls, the Ti/Tv ratio, a critical metric for assessing the amount of false positive SNPs present in the data, was used. The expected Ti/Tv ratio for WES data in the literature varies between 3.0-3.3[75]. A combination of residual false positive makes the Ti/Tv ratio lower at new sites when compared to known sites.

The Ti/Tv ratios for the six samples were calculated using VCFtools (version 0.1.12a) for each of the four variant calling tools and for the combined analysis in order to decide the approach that has a lower rate of false positive SNPs.

The GATK VariantRecalibrator tool was used to evaluate the probability that each call is real. This tool uses known sites (from databases such as dbSNP, Mills and 1000 GP gold standard and HapMap) to estimate the relationship between SNP call annotations and the probability that a SNP is a true genetic variant versus a sequencing error or data processing artifact[76].

Lastly, all the variants in the call set were QC evaluated based on seven annotations parameters (Quality by Depth [QD], MAPping Quality [MAPQ], read DePth [DP], Fisher Strand [FS], Mapping Quality Rank Sum Test [MQRankSum], Haplotype Score and ReadPosRankSum). Briefly:

1) QD is another quality parameter calculated by GATK, which indicates the confidence in a variant (based on the QUAL score), at a given site, over the coverage of

the reads that do not match a given site before being filtered. This parameter is normalized by read length (the higher the length, the more confident is the QD score). Low scores (below 10) are indicative of false positives calls and artifacts[77].

2) MAPQ is a quality parameter calculated by alignment tools (in our case, BWA) that measures the quality of the alignment for the sequencing reads[78]. BWA aligns the sequencing reads with a reference genome (we used hg19) and confers a score between 0 and 60, in which 0 indicates a low mapping quality (less confidence in the alignment) and 60 represents the maximum confidence in the alignment (all the bases are correctly mapped with the reference)[60]. When no cut-off is available for the MAPQ score it is recommended to draw the distribution of mapping quality scores and examine this distribution of outliers[79]. The author of BWA introduced the Base Alignment Quality (BAQ) score, based in the Hidden Markov Model, and implemented it into Samtools as a default parameter. Since InDels often lead to alignment artifacts BAQ score decrease the base quality scores for bases around insertion and deletion events in the sequence reducing the false positive SNP calls. However, a study lead by Guo *et al.* has shown that if BAQ and GATK's local realignment are used consecutively, this increases the false positive SNP calls[79]. This is due to the fact that both tools intend to correct false SNPs caused by insertion and deletions, so, when applying this two tools consecutively will cause an over-correction[79].

3) Coverage or read depth (DP) is a quality parameter that indicates the number of times that each read was sequenced. However, this could be skewed easily by the high-depth regions during exome sequencing, in a phenomenon called unspecific biding where certain regions of the genome have much higher depth than usual. In order to obtain a more realistic value, we used the DP parameter calculated by GATK, that describes the total depth of the reads, for each variant, that passed the caller's internal quality control metrics, like MAPQ > 10, for example[80]. For our data, the range for this parameter varies between 4 and 255.

4) FS is a quality parameter that detects the strand bias in the reads, which means it will identify the number of reference and/or alternative alleles were seen on the forward and reverse strand. This score is calculated with the Fisher's exact test to obtain the p-value and calculate the Phred-scale score ($\mathrm{FS} = -10\log_{10}(\mathrm{p-value})$). The p-value is calculated creating a contingency table (2 x 2) with the number of strands (negative and positive) for the reference and alternate alleles. Higher scores are indicative of false positive calls. The recommended value for this parameter for SNPs

and InDels variants are below 60 and 200, respectively. One example of the Fisher Strand bias calculation is described in Appendix B.

5) MQRankSum is a quality parameter that measures the mapping quality of heterozygous calls (reads with reference allele *versus* alternate allele). It uses the Wilcoxon Rank Sum test to test if two samples are or not the same. The score indicates how many times the standard deviation is distant from the mean. The recommended value for this parameter for SNPs variants are above -12.5.

6) Haplotype Score is a quality parameter that measures the consistency of the site with two segregating haplotypes. This score is calculated for each consensus haplotypes against one of the consensus haplotypes found in a prior set, for each read. The haplotype score is the mean of haplotype scores of all reads for each locus (in a window of 21 bp). It generates one or two consensus haplotypes with the best quality scores (that is the lowest score). The recommended value for this parameter for SNPs variants is below 13.

7) ReadPosRankSum is a quality parameter that measures the distance from the end of the read, for reads with the alternate allele. If the alternate allele is only seen near the extremities of the read it is indicative of an error. It uses the Wilcoxon Rank Sum test to test if two samples are or not the same. The score indicate how many times the standard deviation is distant from the mean. The recommended value for this parameter for SNPs and InDels variants are above -8 and -20, respectively.

For SNPs, all variants with a QD < 2.0, FS > 60.0, MAPQ < 40.0, HaplotypeScore > 13.0, MQRankSum < -12.5, ReadPosRankSum < -8.0, and DP < 4 were filtered out. For InDels variants, all variants with a QD < 2.0, FS > 200.0, ReadPosRankSum < -20.0, and a DP < 4 were filtered out.

In addition, all variants present in the X, Y and mitochondrial chromosomes were filtered out. At the end of all these steps of the variant discovery process twelve VCF (variant call format) files were obtained per individual sequenced: six VCF files for SNP variants and six VCF files for InDel variants were created for each of the six family members under study (III.2-080004, III.4-080001, III.6-080005, IV.1-080002, IV.3-090095 and IV.6-090044). Moreover, QC is present in the whole pipeline.

3.5.1.4 Variant annotation

Many bioinformatics tools are available to functionally annotate genetic variants detected in the human genome as well as their effect prediction. Examples of these are ANNOVAR (functional annotation of genetic variants from high-throughput sequencing data), SNPSift (http://snpeff.sourceforge.net/SnpSift.html - version 3.6b; a collection of tools to manipulate VCF files that is part of SNPeff) and VEP ensembl (Variant Effect Predictor - http://www.ensembl.org/info/docs/tools/vep/index.html - version 75). The last two tools were used to make the annotation of the final variant results that had a *Q* above 50. Both tools require an input file containing the chromosome, start/end positions, reference nucleotide and observed nucleotides for a given a list of variants.

SNPSift was used with three specific aims: identify variants that are present in databases, predict whether an amino acid substitution affects the protein function, and give a conservation score.

VEP ensembl was used to make gene-based annotation (identifies if single nucleotide variants (SNVs) cause protein coding changes and which are the amino acids affected), and region-based annotations (which identifies variants in specific genes/genomic regions). If there are several transcripts in a specific position, VEP choses the variant by the canonical, biotype status and length of the transcript (along with the ranking of the consequence type per variation).

dbSNP 141 (http://www.ncbi.nlm.nih.gov/SNP/) and 1000 GP (http://www.1000genomes.org/) were used to obtain the SNP identification and allele frequencies, respectively. The dbSNP single-nucleotide polymorphism database is a repository for common SNPs and small InDels[36]. The 1000 Genomes Project is a public reference database for DNA polymorphisms which has a 95% completeness of variants allele frequency[81] We also used two functional prediction tools (SIFT and PolyPhen-2) and one conservation tool (GERP++).

SIFT (Sorting Tolerant From Intolerant - http://sift.jcvi.org/) and PolyPhen-2 (http://genetics.bwh.harvard.edu/pph2/) integrate and predict the effect of coding NSVs on protein function, based on the degree of conservation of amino acid residues in sequence alignments derived from closely related sequences.

SIFT scores range between 0 and 1, and scores below 0.05 are predicted to affect protein function[82]. This tool assumes that important positions in a protein sequence have

been conserved through evolution and therefore substitutions at these positions may greatly affect protein function, thus assessing the effect of the amino acid substitution[82].

Polyphen-2 uses the characterization of the substitution site as a feature, comparing the ancestral (reference) allele with the mutant (alternate) allele[36,83]. In this tool a variant is classified as "probably damaging" if its probabilistic score is above 0.85, as "possibly damaging" if its probabilistic score is above 0.15, and the remaining variants are classified as benign[83]. Also, PolyPhen-2 maps the non-synonymous SNPs with a known 3D structure, using the Dictionary of Protein Secondary Structure (DSSP) database to obtain informative features of the protein structure and to see if the substitution is likely to destroy important protein features, such as the hydrophobic membrane[36,83].

Regions that remain conserved over large evolutionary time scales are likely to be involved in key biological processes, unlike the less conserved sites, which accumulate more mutations[84]. To assess the level of constraint (sites that show fewer substitutions than would be expected to occur with neutral evolution) at a site or region of the genome, GERP++ (Genomic Evolutionary Rate Profiling) software tool was used. This tool aligns the DNA sequences of many divergent species and identifies sites under evolutionary constraint, aggregating these sites into longer, potentially functional sequences[84,85]. GERP++ scores vary between -12.36 and 6.18 - the higher the score, the more conserved is the site. This score is based on the alignment of 35 mammalian species sequences[86].

3.6 SNP selection and validation

For filtering and scoring the variants and detection of disease-causing mutations, there are several possible strategies depending on the mode of inheritance and on the number of affected/non-affected individuals sequenced. Sequencing several affected and non-affected individuals from the same family dramatically improves the filtering process. Potentially functional variations (NSVs and/or variants with a probable splice site effect) present in all affected individuals and absent in the unaffected relative(s) were prioritized. In the identification of disease-related alleles, the main challenge is to identify the causal alleles among the background of non-pathogenic polymorphisms and sequencing errors[56]. Several strategies for allele discovery using NGS also rely on the mode of inheritance of a trait, the pedigree, population structure, whether it is a *de novo*

or inherited variant, and the extent of *locus* heterogeneity for the trait under study. Moreover, the sample size needed to provide adequate power to detect trait-associated alleles, and the selection of the most successful analytical framework are influenced by these factors[56].

Data was divided in two main categories: SNPs and InDels. These categories contain several types of alterations in the coding DNA sequence (CDS) region. The SNP category includes non-synonymous, synonymous, stop-gain and stop-loss variants. The InDel category encompasses frameshift insertions/deletions, non-frameshift insertions/deletions, stop-gain and stop-loss variants. Variants in the CDS region were the focus of this research because this is the portion of DNA that encodes for proteins. For SNPs, NSVs were analyzed first (because they lead to different amino acids that could be damaging for protein function, thus a probable cause for the disease) and also stop-gain/stop-loss variants. For InDels, were analyzed frameshift insertions/deletions, non-frameshift insertions/deletions, frameshift/non-frameshift block substitution and stop-gain/stop-loss variants. For both SNPs and InDels, the search was expanded for other regulatory regions (3'UTR, 5'UTR, upstream, downstream, intronic, intergenic and regulatory) that were identified but did not validate.

For the filtering process were used two different filters. The first filter applied to the variants under study aims to identify variants that are absent from the unaffected relative and present in all affected individuals, to highlight novel alleles shared only among affected family members.

Next, assuming that the PDB causative variant in the family under study is novel and not described in public databases, variants present in dbSNP141 and 1000 GP databases were filtered out.

SIFT and Polyphen-2 scores were used to assess if the variants are possibly damaging to the protein function[87].

After the filtering steps, validation of the selected mutations of interest was performed using traditional Sanger sequencing at the Sequencing Unit of *Instituto Gulbenkian de Ciência* (IGC).

The standard nomenclature recommendations of the HGVS (Human Genome Variation Society) were followed to name variants. Standard mutation nomenclature based on coding DNA reference sequences requires prefixes "c." and numbering starts with number 1 for the first nucleotide in the sequence[88].

3.7 Primer design & preparation

PCR and sequencing primers were designed using Primer3 (http://bioinfo.ut.ee/primer3-0.4.0/primer3/)[89]. This programme considers possible combinations for reverse and forward primers to look for the best primer pairs[89]. Primer3 takes into consideration parameters such as sequence specificity, similar melting temperatures (Tm) of primer pairs, low GC content and pairing of primers with low probability of forming loops[89]. Primers were on average 20 nucleotides long for specific targeting and amplification, and spanning a specific region of interest (Table 4).

To confirm the uniqueness of each primer sequence Blast-Like Alignment Tool (BLAT) program (http://genome.ucsc.edu/cgi-bin/hgBlat?command=start) was used, a local alignment tool from University of California–Santa Cruz (UCSC) database. This tool compares each DNA sequence against the whole human genome and scores it giving a percentage of identity. Also, the uniqueness of the forward and reverse primer for each region that wants to amplify was confirmed via *in silico* PCR (http://genome.ucsc.edu/cgi-bin/hgPcr?command=start), which searches how specific each pair of primers is, and gives the PCR product and its characteristics.

Lyophilized primers were re-suspended in MilliQ water, making final 100 μM stock solutions. In turn, these were diluted to 8.27 μM (working solution) and stored at -20 ^{o}C.

Table 4. Primers used to Sanger sequence candidate mutations and quality control variants.

CDS change	Gene	Chr	Ref allele	Obs allele	Primer sequence (5' --> 3')	Primer length (bp)	PCR fragment (bp)
c.G180A	*SERINC2*	1	G	A	TATATGACCCAGCCTCCCTCT (f)	21	460
					AGATCATCAGTGCACCCAAAC (r)	21	
c.2163_2168del	*PLEKHG5*	1	TTCCTCC	T	CCCACAGTGTTCATGACAAGAG (f)	22	379
					AGATTAGGGAGATGCTGGTCAC (r)	22	
c.C2264T	*NUP210*	3	G	A	CACACTCACCACCTGCTTGT (f)	20	204
					CTGTTCACTGTGCCCTACCA (r)	20	
c.C871T	*MLL3*	7	G	A	TTACAACATTTGTTATTTTC (f)	20	332
					TACTTGTGATATACAGAGAGT (r)	21	
c.C4786T	*KIAA1875*	8	C	T	GATGAGACTGAGGGGTGAGTG (f)	21	394
					CCTGCAGAAAGAAATTCTCTGG (r)	22	
c.C8800T	*CUBN*	10	G	A	AGTCAAGAGGACCACTGACAGA (f)	22	152
					TGATGACTTTTTGTTCCCACA (r)	21	
c.T1933C	*PML*	15	T	C	CCAAGGTGAGGTCTCTAGATGG (f)	22	350
					GGAATTCCCACAGCCTGTTAAT (r)	22	
c.C53T	*NLRC3*	16	G	A	AGACAGCTTCTTGGAGTCTCGT (f)	22	444
					GACAGGAAGGAAGAATGAGGTG (r)	22	
c.T566C	*SRL*	16	A	G	TCTCCCAAAGAGTTGGGATC (f)	20	417
					CACCAGGCATATACACATGCTT (r)	22	
c.C478G	*EMR2*	19	G	C	CTGCTTTGGAGGACCTGACT (f)	20	150
					AACACCCTCGGCAGCTACAC (r)	20	

Chr.: Chromosome; CDS change: Coding DNA sequence change; Ref.: Reference; Obs.: Observed; PCR: Polymerase chain reaction; bp: Base pair; f: Forward; r: Reverse.

3.8 Polymerase Chain Reaction (PCR)

The first step in Sanger sequencing is the amplification of the regions of interest by the PCR method. PCR is an enzymatic process in which a specific DNA region is replicated over and over again to yield many copies of a particular sequence of interest. This process involves heating and cooling samples in a precise thermal cycling pattern over ~30 cycles. During each cycle, a copy of the target DNA sequence is generated for every molecule containing the target sequence. The oligonucleotide primers, that are complementary to the 5′- and 3′-ends of the sequence of interest, define the boundaries of the amplified product. Theoretically, after 30 cycles, approximately one billion copies of the target region (DNA template) have been generated. This PCR product ('amplicon') is then in sufficient quantity to be easily quantified and verified by a variety of techniques[90].

The PCR reaction requires a genomic template, deoxyribonucleotide triphosphates (dNTPs), magnesium chloride ($MgCl_2$), primers, buffer and a polymerase enzyme. The PCR reaction typically contained final concentrations of 0.2 mM of each dNTP, 0.625 U/μL AmpliTaq Gold, 1.5 mM $MgCl_2$, 1x AmpliTaq Gold buffer, 0.5 mM of each primer (forward and reverse) and 10 ng/μL of template DNA. All these reagents were supplied by NZYTech, apart from the primers that were from Invitrogen and the DNA. The final PCR volume was 12.5 μL (Table 5). The PCR reaction was performed on a PCR 2720 Thermal Cycler (Applied Biosystems).

Table 5. Components for the PCR reaction.

Component	Volume in a 12.5 μL reaction (μL)
Water	2.25
NZYTaq Colourless Master Mix (2 x)	6.25
Forward primer (0.5 mM)	0.50
Reverse primer (0.5 mM)	0.50
DNA (10 ng/μL)	3.00

Generally a PCR cycle includes three steps:

1. Denaturation at 94^{o}C (the double DNA strand melts to single stranded DNA).

2. Annealing at approximately 60^{o}C (the primers bind specifically to their complementary sequence in the single stranded DNA; the annealing temperature usually depends on the Tm of the expected duplex).

3. DNA synthesis (extension of the complementary strand is initiated by the annealed primer and occurs most efficiently at 72^{o}C).

Each cycle is repeated 25-35 times and is finished by a longer final extension at 72^{o}C to complete all synthesis of the amplified region.

A PCR *stepdown* program was used to amplify the candidate variants because this is a robust method that prevents primers to amplify non-specific regions. PCR initiates with an annealing temperature above the optimum annealing temperature, which decreases in each cycle, helping to ensure a competitive advantage for the right target sequence. Thus, in the initial cycles, the PCR primers will mostly hybridize with the region of interest, increasing the specificity substantially. Then, in subsequent cycles, the temperature decreases, diminishing the specificity, while increasing the amplification efficiency[91,92].

Optimization of PCR conditions was carried out by changing/adjusting the annealing temperature and/or number of cycles, as appropriate, until the optimal conditions were achieved. When the successful amplification of the targeted region was obtained, 6 μL of product was analyzed on an agarose gel by electrophoresis for product size confirmation (as previously described in sub-chapter 3.3). The conditions that provided optimal results were then used to amplify the remaining samples, and are specified for each primer pair in Appendix D. When the amplification was suboptimal, further systematic optimization of the PCR conditions was attempted. The KAPA2G Robust HotStart (Kapa Biosystems, Woburn, USA) was used in four variants (c.C4786T, c.C53T, c.G180A and c.2163_2168del). This kit is suited for the amplification of DNA templates with a high GC or AT content. It contains a DNA polymerase combined with a proprietary antibody that inactivates the enzyme until the first denaturation step, which eliminates spurious amplification products resulting from non-specific priming events during reaction setup and initiation, and increases overall reaction efficiency[93]. KAPA Enhancer 1 improves the amplification of difficult

templates by increasing primer specificity. KAPA2G GC Buffer helps the amplification of sequences with high GC content (Table 6).

Table 6. PCR reaction using KAPA2G Robust HotStart kit.

Component	Volume in a 25 µL reaction (µL)
Water	5.30
KAPA2G GC buffer (5 x)	5.00
MgCl2 (25 mM)	0.50
KAPA Enhancer 1 (5 x)	5.00
dNTP mix (10 mM)	0.50
Forward primer (10 µM)	1.25
Reverse primer (10 µM)	1.25
DNA (90 ng/ µL)	6.00
KAPA2G Robust Hotstart DNA polymerase (5 units/ µL)	0.20

3.9 Sanger sequencing

After the amplification of regions of interest, traditional Sanger sequencing was performed to validate the candidate variants in family 1 and 2. This process involves the incorporation of dideoxyribonucleotide triphosphates (ddNTPs) as chain terminators followed by a separation step capable of single nucleotide resolution. There is no hydroxyl group at the 3′-end of the DNA nucleotide with a ddNTP and therefore chain growth terminates when the polymerase incorporates a ddNTP into the synthesized strand. Extendable dNTPs and ddNTP terminators are both present in the reaction mix so that some portions of the DNA molecules are extended. At the end of the sequencing reaction a series of molecules are present that differ by one base from one another[90].

Each DNA strand is sequenced in separate reactions with a single primer. Either the forward or reverse PCR primers are used for this purpose. Four different colored fluorescent dyes are attached to each four ddNTP. Thus, ddTTP (thymine) is labeled with a red dye, ddCTP (cytosine) is labeled with a blue dye, ddATP (adenine) is labeled with a green dye, and ddGTP (guanine) is labeled with a yellow dye although it is typically displayed in black for easier visualization. These are similar dyes for Short Tandem Repeat (STR) polymorphisms detection. Fluorescent dye labels have simplified DNA sequencing as have the widespread use of automated detection systems and capillary electrophoresis[90].

After obtaining successful PCR products and confirming them on a 2% agarose gel, products are prepared for the sequencing reaction and purified prior entering the sequencing platform. The Cycle Sequencing BigDye Terminator v1.1 protocol was used. Briefly, to each 1 μL PCR product (90 ng/μL) was added 4 μL of MilliQ water, 2 μL of buffer, 2 μL of Terminator Ready Mix Dye and 1 μL of primer (3.2 pmol). The sequencing reactions were performed in the 2720 Thermal Cycler (Applied Biosystems) using caps and heated lids, with the following conditions:

96ºC for 1 min

25 cycles {
- 96ºC for 10 sec
- 50ºC for 5 sec
- 60ºC for 1.15 min
}

Rapid thermal ramp (1ºC/sec) to 4ºC and hold until ready to purify

It was added to each tube 10 μL of MilliQ water, 2 μL of 3M sodium acetate (AcNa, pH 4.6), 50 μL of 95% ethanol and 10 μL of the PCR product obtained. The samples were then centrifuged at maximum speed (13.200 rpm) for 30 min (4ºC), the supernatant carefully discarded and 250 μL of 70% ethanol added. Samples were briefly vortexed and centrifuged at maximum speed (13.200 rpm) for 15 min (4ºC). Lastly, the pellet was dried at room temperature and then was frozen until being shipped for IGC to process.

3.10 Sequence analysis

Sequence files generated at IGC were imported to the Staden package (Pregap4, Trev and Gap4)[94] and checked individually for variation by comparison with a reference sequence.

Pregap4 prepare trace data for assembly, such as trace format conversion and quality analysis. The configuration modules selected for analysis are 1) estimation of base accuracy (estimates a confidence value for the base called), 2) initialize experiment files (for the sequence assembly), 3) quality clip (to determine the regions where the sequence has low quality to use for reliable assembly), 4) interactive clipping (to "call"

the Trev program to see the chromatogram files) and 5) Gap4 shotgun assembly (assembles the processed sequences into Gap4 using its own assembly engine)[94].

Trev exhibits the original sequence, the confidence value for each base call, and allows editing the chromatogram data prior to assembly into a Gap4 database.

Gap4 mainly carries out the sequence assembly and checking, contig ordering based on read pair data, and editing. It provides a graphical view of the contigs, readings and traces. Both primer pairs were aligned and compared to the reference and the position of the relevant variants and/or novel SNPs/InDels was highlighted. This programme also provides a confidence measure for each allele call per position.

3.11 Reagents and buffers

- ✓ MilliQ water
- ✓ Isopropanol
- ✓ EtOH 70% and 95%
- ✓ RBC Lysis Buffer (RBC Bioscience Corp., New Taipei City, Taiwan)
- ✓ Cell Lysis Buffer (RBC Bioscience Corp., New Taipei City, Taiwan)
- ✓ Protein Remove Buffer (RBC Bioscience Corp., New Taipei City, Taiwan)
- ✓ NZYTaq 2x Colourless Master Mix (NZYTech, Lisboa, Portugal)
- ✓ TAE 1x
- ✓ Agarose powder (NZYTech, Lisboa, Portugal)
- ✓ Loading dye (NZYTech, Lisboa, Portugal)
- ✓ GreenSafe Premium (NZYTech, Lisboa, Portugal)
- ✓ 1x NZYDNA Ladder VI (NZYTech, Lisboa, Portugal)
- ✓ KAPA2G Robust HotStart (Kapa Biosystems, Woburn, USA)
- ✓ AcNa 3M (Sigma-Aldrich, inc, St. Louis, USA)
- ✓ BigDye (BigDye Terminator v1.1 Cycle Sequencing Kit, Applied Biosystems by Life Technologies, California, USA)

4. Results

4.1 Family collection

The study was carried out using two Portuguese multiplex PDB families from *Alentejo*. Multiplex families are families with multiple affected individuals. Both families were acertained at *Instituto Português de Reumatologia* (IPR) by Doctor José Vaz Patto and Doctor Filipe Barcelos, two rheumatologists with a special interest in PDB.

All participants received an explanation of the study and signed an appropriate informed consent, ensuring their voluntary participation in our study. Blood samples of eight individuals from the family 1 (Figure 6) and six individuals from the family 2 were collected (Figure 7). The participant's clinical and demographic information was collected using a standardized questionnaire and the data was subsequently stored in an in-house secure database (BCgene - https://bcgene.igc.gulbenkian.pt/bcos/index.html).

This study has been approved by the ethics committee of IPR (Appendix E).

4.1.1 Characterization of the PDB family

The genogram of families 1 and 2 are depicted in Figure 6 and 7, respectively.

Diagnostic tests have been made, in particular the total serum alkaline phosphatase measurement, which is an important biomarker aiding the physician in establishing the diagnosis of PDB. The medical information of the families obtained is summarized in Tables 7 (family 1) and 8 (family 2). At the end of the validation analysis we did not know if the individual ID IV.3-090095 (from family 1) was affected or not.

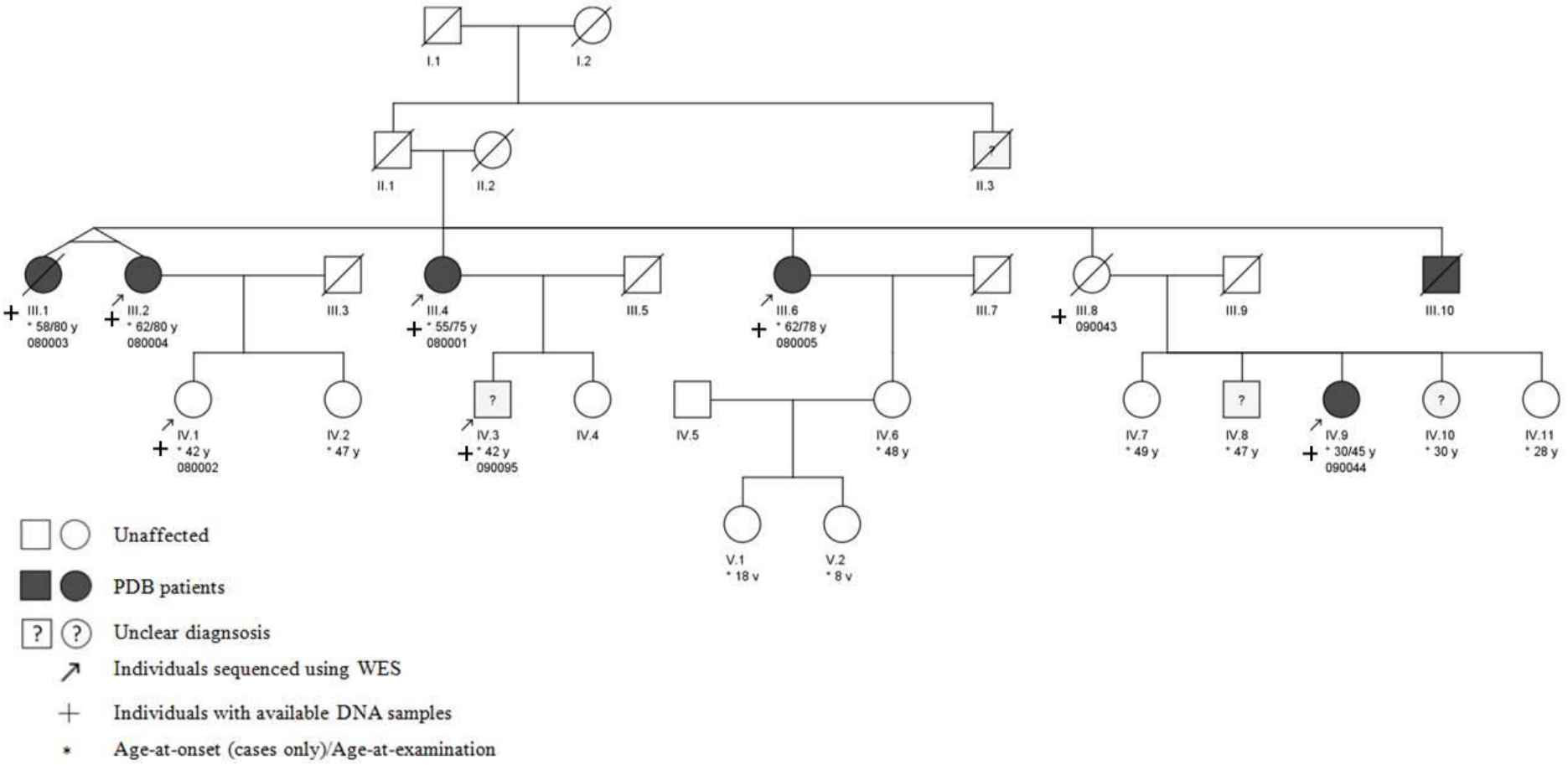


Figure 6. Genogram of the PDB family 1. This genogram was constructed using CeGat Pedigree Chart Designer (version 2.1). The arrows indicate the individuals selected for WES (III.2-080004, III.4-080001, III.6-080005, IV.1-080002, IV.3-090095 and IV.9-090044).

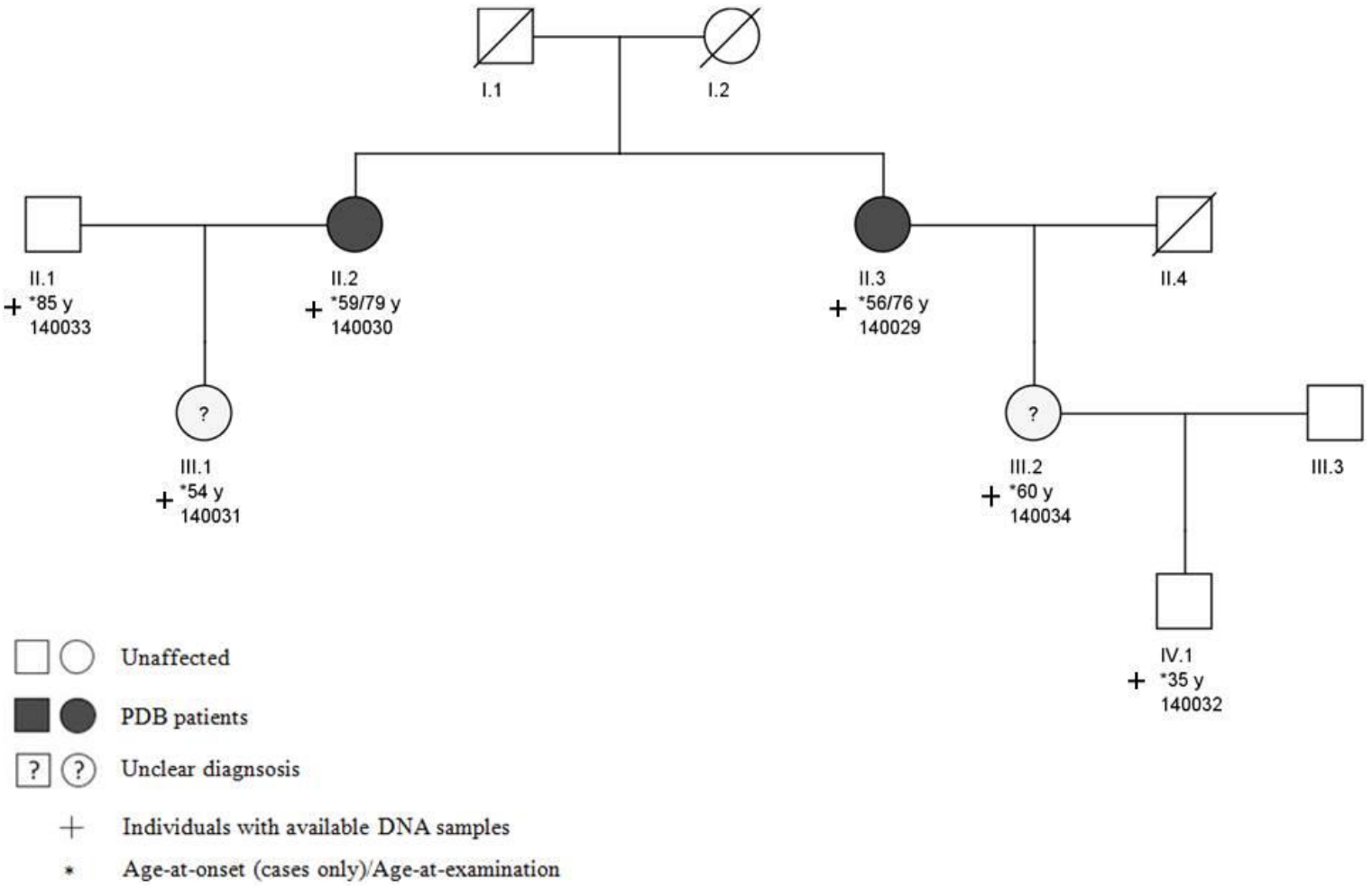


Figure 7. Genogram of the PDB family 2. This genogram was constructed using CeGat Pedigree Chart Designer (version 2.1). All individuals from whom DNA were available (II.1-140033, II.2-140030, II.3-140029, III.1-140031, III.2-140034 and IV.1-140032) were used for validation.

Table 7. Clinical and demographic characterization of the Portuguese multiplex PDB family 1.

Family relative ID	Gender	Year of birth	Affection status	Age at onset	Nº of affected bones	Localization of lesions	Symptoms
III.1-080003	F	1928	Patient	58	12	Cranium, scapula (r), sacrum, pelvis (r), both femurs, patela (r), both tibia, fibula (r)	Skeletal pain, skeletal deformities, fractures (fissure fracture, chalk-stick fracture), hearing deficit, generalized atherosclerosis
III.2-080004	F	1928	Patient	62	9	Cranium, humerus (l), pelvis (l), both femurs, patella (l), tibia (l), both fibulas	Skeletal pain, skeletal deformities, secondary osteoarthritis, hearing deficit, generalized atherosclerosis
III.4-080001	F	1933	Patient	55	7	Cranium, sacrum, pelvis (l), femur (l), both tibia, tarsus (r)	Skeletal pain, skeletal deformities, fractures (fissure fracture, chalk-stick fracture), secondary osteoarthritis, hydrocephalus, slight pericardial effusion
III.6-080005	F	1931	Patient	62	3	Cranium, clavicle (r), pelvis (r)	Skeletal pain, secondary osteoarthritis, hearing deficit
III.8-090043	F	NA	Unaffected	-	-	-	-
IV.1-080002	F	1965	Unaffected	-	-	-	-
IV.3-090095	M	1954	Unclear	52	2	Femur (r), tibia (l)	Skeletal pain
IV.9-090044	F	1963	Patient	30	4	Cranium, sacrum, humerus (r), pelvis (r)	Skeletal pain, secondary osteoarthritis

F: Female; M: Male; l: left; r: right; NA: Not available.

Table 8. Clinical and demographic characterization of the Portuguese multiplex PDB family 2.

Family relative ID	Gender	Year of birth	Affection status	Age at onset	N° of affected bones	Localization of lesions	Symptoms
II.1-140033	M	1928	Unaffected	-	-	-	-
II.2-140030	F	1935	Patient	59	18	Cranium, both clavicles, vertebral column, both humerus, both hand carpals, both hand metacarpals, both hand phalanges pelvis (r), both patella, both foot tarsals, both foot metatarsals, both foot phalanges	Skeletal pain, skeletal deformities, fractures (fissure fracture, chalk-stick fracture), secondary osteoarthritis, vascular cerebral accident, hypertension, hyperparathyroidism, lupus, rheumatoid arthritis
II.3-140029	F	1937	Patient	56	9	Cranium, both clavicles, both humerus, both pelvis, patella (r), foot phalanges (r)	Skeletal pain, skeletal deformities, fractures, secondary osteoarthritis, hearing deficit, generalized atherosclerosis, hyperparathyroidism
III.1-140031	F	1960	Unclear	-	3	Vertebral column, both hands phalanges	Skeletal pain, secondary osteoarthritis, hypertension, gouty diathesis
III.2-140034	F	1953	Unclear	-	8	Both humerus, both hand phalanges, both patella, both foot phalanges	Skeletal pain, secondary osteoarthritis, hyperparathyroidism, diabetes
IV.1-140032	M	1979	Unaffected	-	-	-	-

F: Female; M: Male; l: left; r: right.

4.1.1.1 Sample selection criteria for the WES assay

The use of pedigree information can substantially narrow the genomic search space for candidate causal alleles. Depending on the frequency of the disease-causing allele and the nature of the relationship between the individuals within the family under study one chooses the most informative individuals to further analyze[56].

Sequencing multiple affected individuals from within one family to identify genes with novel variants in a shared region of the exome is the most efficient strategy in our case, since we can substantially reduce the exonic search space by choosing the relatives that share half or less than half of the genetic information between them. Thus, we can reduce the genetic information that is shared between the family relatives increasing the probability of find the risk variants. Siblings from the third generation (III.2-080004, III.4-080001, III.6-080005) were selected since in the case of rare alleles the probability of identity-by-descent given identity-by-state is high, even among distantly related individuals[56]. This is also the case for first cousins that were also selected to study (IV.1-080002, IV.3-090095 and IV.9-090044) in order to substantially restrict our exonic search.

The DNAs from the six selected individuals from family 1 were therefore sent to BGI for WES. Few weeks later, we received from BGI the raw sequencing data as well as their bioinformatics analysis.

4.2 Quality control of the WES data

To validate the WES data and bioinformatics analysis from BGI, four variants with varying quality scores were selected for validation by Sanger sequencing. To ensure that a good concentration of DNA was obtained to proceed to the WES analysis and validation (by Sanger sequencing) all the DNA samples were quantified and run in an agarose gel to confirm DNA quality.

Moreover, three samples (III.4-080001, III.6-080005 and IV.3-090095) were run in the bioanalyzer to make sure that the DNA had high quality for the WES assay, which is a more sensitive technique. Only three of the six samples (sent to BGI) were run in the bioanalyzer because this was the first time that this analysis was performed and for that reason it was necessary to run positive controls from the lab (DNA samples recently extracted), in order to be sure that we obtained optimal results. This is a crucial

step because the WES assay is sensitive to slightly degraded DNA, so it was needed to ensure that our samples (which have been stored for five years) had a similar quality of DNA samples recently extracted.

All 14 samples had an excellent quality and their concentration was above the required (> 40 ng/μL) for the WES and validation assays. In the bioanalyzer gel (Figure F.1. A, Appendix F), none of the three PDB samples seemed degraded unlike the sample 090022 (a control sample of the lab), which had a band at 7000 bp indicating that the DNA was slightly degraded. Also, there is no peak between the two markers (Figure F.1. B, Appendix F) indicating that the genomic DNA of the three PDB samples is not fragmented.

Six quality parameters were used to check BGI's SNPs and InDels results: (1) MAPQ (MAPping Quality), (2) GQ (Genotype Quality), (3) DP (read DePth), (4) AD (Depth Allele by sample), (5) QUAL (QUALity score) and (6) QD (Quality by Depth). Briefly:

1) MAPQ is a quality parameter that measures the quality of the base alignment. The highest score is 60, which indicate that the read matches rightly with the reference genome.

2) GQ (or Phred-scale confidence) is another quality parameter calculated by GATK that indicates if the inferred genotype for each variant is the real genotype, this means, if they are estimated with high/low confidence level[95]. GQ ranges between 0 and 100%.

3) DP is a quality parameter that indicates for each position how many times each variant passed the internal quality control (such as a MAPQ > 10)[60]. A base with a DP of four indicates that this base was read four fold. This first three quality parameters are described in more detail in the sub-chapter 3.5.1.3.

4) AD value is the number of reference (ref) and alternate (alt) alleles observed in each variant read. The total sum of ref and alt alleles is then equal to the DP value calculated, providing complementary information on our data[96].

5) QUAL is a quality score that indicates the probability of a variant (ref/alt) exists at a certain site, given sequenced data (Table 2). This score grows with the amount of NGS data analyzed in the variant calling step[95].

6) Finally, QD is a quality parameter that indicates the variant's confidence at a given site. Low scores are indicative of false positive calls.

To show an example of the interpretation of all six quality parameters, an example for variant c.C1165A is shown in Table 9.

The table 9 indicates that:

- the quality score (QUAL) is 802.92, indicating approximately 1 in 10^{80} chances of error;

- the MAPping Quality (MAPQ) is 60, indicating a good confidence in the alignment;

- the Quality by Depth (QD) is 13.16 (QD above 10 is consider a variant with a good confidence call);

- the genotype value (GT) is 0/1, which means this individual is heterozygous at this *locus*;

- the depth allele by sample (AD) shows a total number of reference nucleotide (G) is 33 and the alternate nucleotide (T) is 28;

- the read DePth (DP) is 61, which means that this variant as a coverage of 61 fold;

- the genotype quality (GQ) is 99%;

- the genotype likelihood (PL) presents three different values who correspond to: 0/0 (homozygous) = 833; 0/1 (heterozygous) = 0; 1/1 (homozygous) = 939, so the genotype is 0/1, because the value of genotype likelihood is equal to zero.

Table 9. Example of WES QC parameters for the non-synonymous variant (NSV) c.C1165A in *CUBN*.

CDS change	Gene	Mutation type	Chr	Start/End (bp)	Ref	Obs	QUAL	MAPQ	QD	GT	AD	DP	GQ	PL
c.C1165A	*CUBN*	NSV	10	17.147.521/ 17.147.521	G	T	802.92	60.00	13.16	0/1	33;28	61.00	99.00	833,0,939

CDS change: Coding DNA sequence change; NSV: Non-synonymous variant; Chr: Chromosome; bp: Base pair; Ref: Reference; Obs: Observed; MAPQ: MAPping Quality; QD: Quality by Depth; GT: GenoType; AD: Depth Allele by sample; DP: read DePth; GQ: Genotype Quality; PL: Genotype Likelihood.

Based on the BGI bioinformatics analysis one variant (c.C8800T) with good quality (MAPQ ~60, and DP of ~36x), two variants (c.C2264T and c.C871T) with a medium quality (one with low MAPQ ~36, and another with a mean DP ~25x), and one variant (c.C478G) with bad quality (low MAPQ ~24.94 and DP ~10x) were selected for quality control (Table 10). For each variant one PDB patient and one control from family 1 were randomly sequenced to confirm BGI's results as either real or false positive/negative.

In addition, for each variant were counted how many reference and/or observed alleles were seen in the reverse and/or forward strand to verify if the alternate allele is disproportionately represented on one strand migh be indicative of a false positive variant (Fisher strand bias – Table 11).

At this point, the PCR for the "bad" quality variant (c.C478G) was not optimized because this is inserted in a repetitive region, and was very difficult to amplify the region of interest (even after several attempts). This is reflected on the BGI results, since the MAPQ attributed to this variant is very low (MAPQ = 24.94), corresponding to a low confidence in mapping the read with the reference genome. The same applies to the DP (DP = 10x), which means that just a few reads were actually counted.

Two PCRs (one from a "good" quality variant – c.C8800T - and another from the variant with "medium" quality – c.C2264T) are represented in Figure 8. The respective sequencing chromatograms are represented in Figures 9B and 10B for one PDB patient and one control.

Afterwards, these four regions were visualized using IGV to enquire if the BGI results corresponded to the results obtained for the "good" (Figure 9), "medium" (Figures 10 and 11) and "bad" quality variants (Figure 12) both in IGV and our Sanger sequencing results.

The results visualized in IGV for the two variants (c.C8800T and c.C2264T) are consistent with the results obtained by BGI (Figure 9A and 10A). However, contrary to the results obtained by BGI, variants c.C871T and c.C478G are also present in the control IV.1-080002 (Figure 11A and 12A).

Table 10. Selected variants for the QC parameters assessment. These variants were chosen mainly following two main quality control parameters (DP and MAPQ) and are organized in three quality categories (Good, medium and bad).

CDS change	Gene	Chr	Start/End (bp)	Ref Allele	Obs Allele	SNP ID	Variants quality	DP	MAPQ	QUAL	AD	QD	GQ
c.C8800T	*CUBN*	10	16.930.521/ 16.930.521	G	A	.	Good	36.00	60.00	493.61	16;20	13.71	99.00
c.C2264T	*NUP210*	3	13.399.786/ 13.399.786	G	A	rs6795271	Medium	25.00	60.00	250.47	15;10	10.02	99.00
c.C871T	*MLL3*	7	151.970.931/ 151.970.931	G	A	rs56850341		84.00	36.98	1204.08	23;50	16.49	99.00
c.C478G	*EMR2*	19	14.877.799/ 14.877.799	G	C	rs12976472	Bad	10.00	24.94	75.48	0;10	7.55	9.03

CDS change: Coding DNA sequence change; Chr: Chromosome; Ref: Reference; Obs: Observed; bp: Base pair.

Note: All variants are exonic NSVs.

Table 11. Strand bias for the QC variants selected. For each variant, was visualized how many reference and/or alternate alleles were seen in the reverse and/or forward strand using IGV and was calculated the Fisher Strand bias (when FS = 0 there is no bias and FS > 0 indicates bias, so the higher the score more bias is present).

		III.4-080001			IV.1-080002			III.2-080004			III.6-080005			IV.9-090044			IV.3-090095		
CDS change	**Strand**	**Ref**	**Obs**	**FS**	**Ref**	**Obs**	**FS**	**Ref**	**Obs**	**FS**	**Ref**	**Obs**	**FS**	**Ref**	**Obs**	**FS**	**Ref**	**Obs**	**FS**
c.C8800T	**Positive**	.	.	.	.	.	.	8	5	0.00	8	6	0.00	5	6	0.00	.	.	.
	Negative	.	.		.	.		18	10		23	17		10	14		.	.	
c.C2264T	**Positive**	16	14	0.00	.	.	.	15	9	3.98	8	18	9.24	10	12	4.63	.	.	.
	Negative	2	2		.	.		0	1		2	0		3	1		.	.	
c.C871T	**Positive**	10	72	11.33	13	45	6.24	19	44	0.00	15	50	0.00	18	49	5.87	22	66	0.95
	Negative	6	13		7	12		4	8		5	19		8	11		7	19	
c.C478G	**Positive**	26	2	8.28	26	2	0.00	0	10	0.00	1	4	3.43	0	7	0.00	.	.	.
	Negative	9	3		9	1		0	1		0	6		0	5		.	.	

CDS change: Coding DNA sequence change; Ref: Reference; Obs: Observed; FS: Fisher Strand.

Note: Individual IV.1-080002 presented variants c.C871T (*MLL3*) and c.C478G (*EMR2*) using IGV in our analysis, but in BGI results they were absent

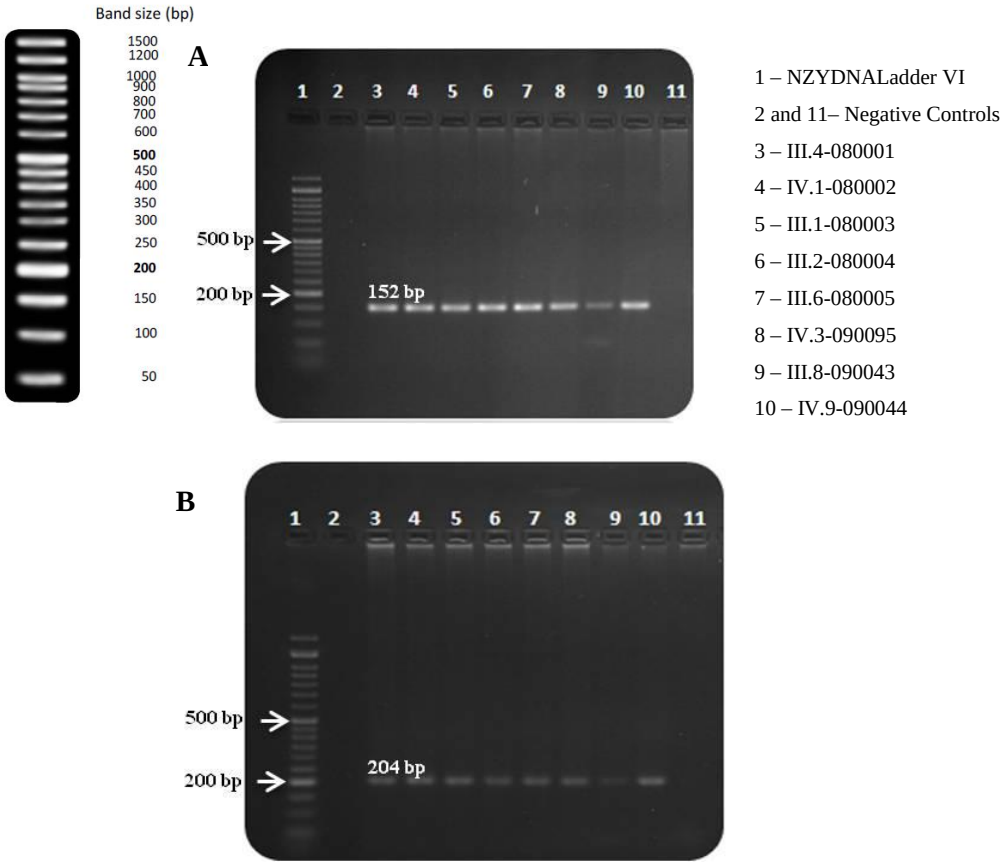
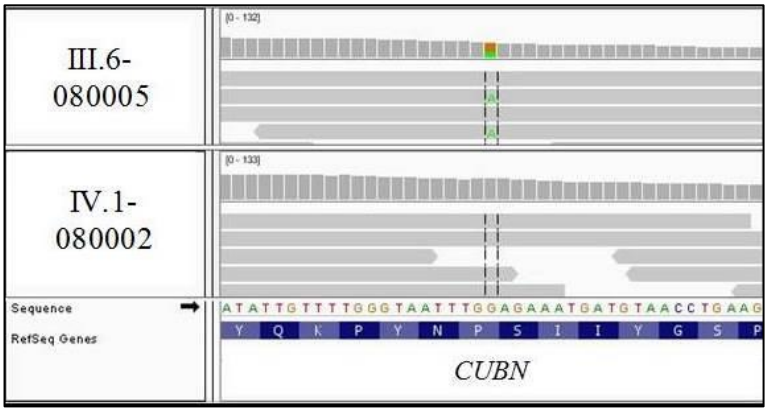


Figure 8. PCR amplification of the c.C8800T (A) and c.C2264T (B) variants in individuals from family 1. The NZYDNALadder VI was loaded in the first lane of a 2% agarose gel and the PCR products are running at the expected location for 152 bp (A) and 204 (B) fragments, respectively.

A



G G G T A A T T T G [G/A] A G A A A T G A T G T A

B

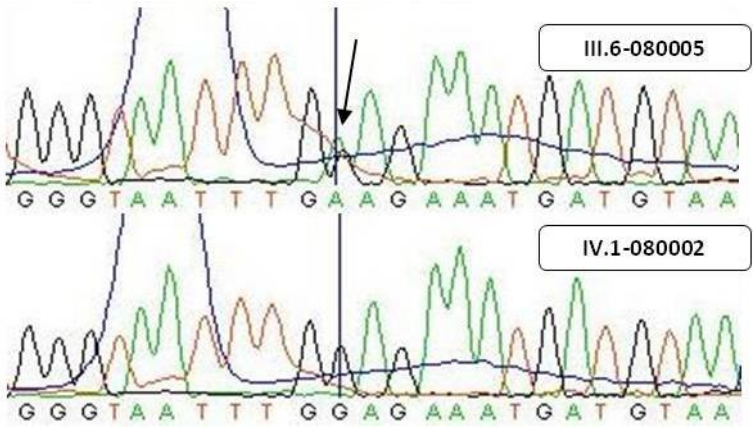
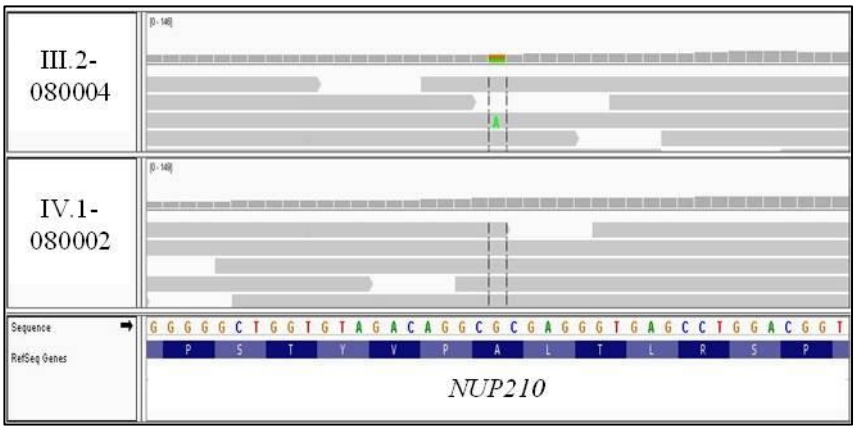


Figure 9. "Good" quality variant (c.C8800T) according to IGV (A) and chromatograms obtained via Sanger sequencing (B). A - c.C8800T variant with the mutated A allele in patient III.6-080005, and absent in the control (IV.1-080002), depicted between the dashed black lines using IGV. The reads are depicted as grey arrows aligned by base and above them there is a histogram that reflects the coverage per base. Polymorphisms are highlighted in the respective read and the deletions are represented as a black line. Also, the reference sequence (hg19) and respective amino acids are also shown below, when possible. **B -** Here one can see the peak for the mutated A allele for c.C8800T variant in patient III.6-080005 (arrow), and absent in the control (IV.1-080002).

A



G A C A G G C [G/A] C G A G G G T G A

B

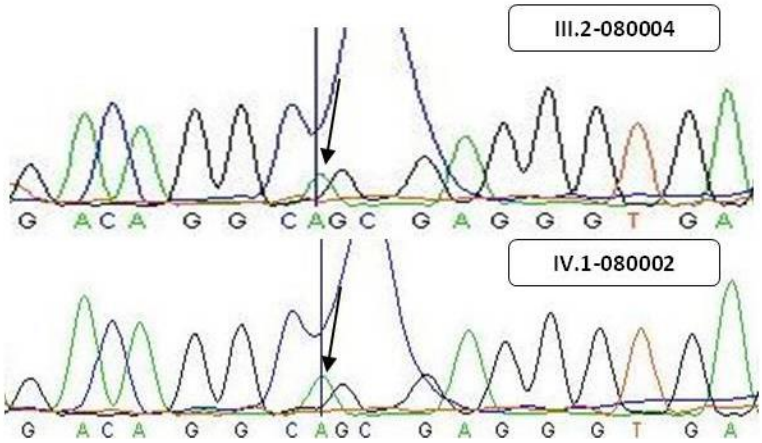
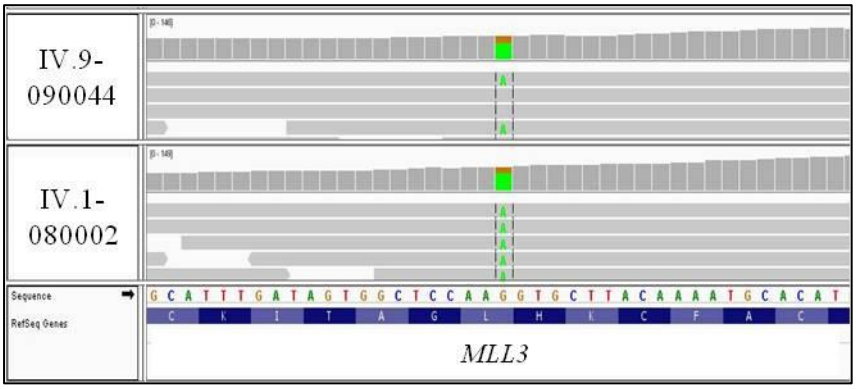


Figure 10. "Medium" quality variant (c.C2264T) according to IGV (A) and chromatograms obtained through Sanger sequencing (B). A - c.C2264T variant with the mutated A allele in patient III.2-080004, and absent in the control (IV.1-080002). **B -** Here one can see the mutated A allele for c.C2264T variant in patient III.2-080004 (above, black arrow). It is also present in the control IV.1-080002 (below, black arrow).

A



G C T C C A A [G/A] G T G C T T

B

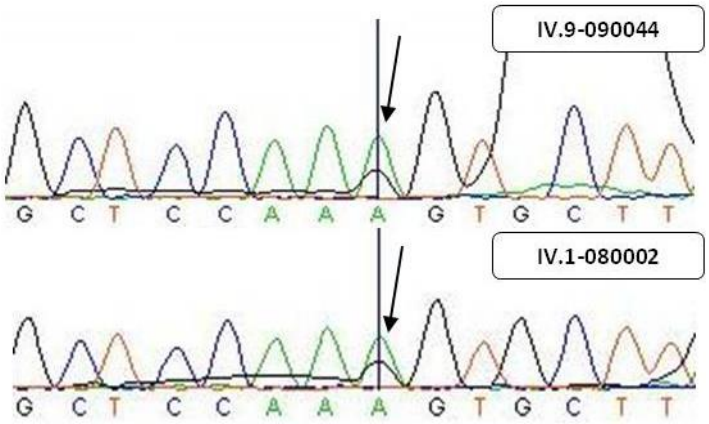


Figure 11. "Medium" quality variant (c.C871T) according to IGV (A) and chromatograms obtained via Sanger sequencing (B). A - The c.C871T variant with the mutated A allele in patient III.9-090044 is here depicted, also present in the control (IV.1-080002). **B -** Here one can see the mutated A allele for the c.C871T variant in patient IV.9-090044 (above, black arrow). This variant is also present in the control IV.1-080002 (below, black arrow).

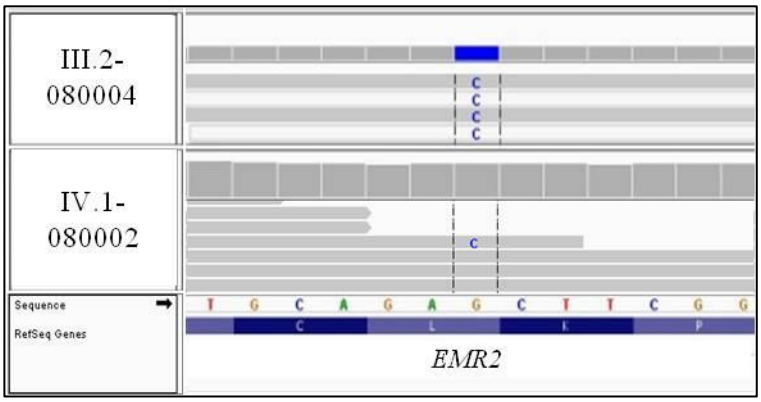


Figure 12. "Bad" quality variant (c.C478G) according to IGV. Here one can see the c.C478G variant with the mutated C allele in patient III.2-080004, also present in the control (IV.1-080002).

A perfect match was obtained between the alleles reported in the BGI WES analysis and the Sanger sequencing performed for the "good" quality variant c.C8800T (Figure 9). As shown in Figure 9, patient III.6-080005 is heterozygous having both the reference (G) and alternate (A) alleles at 16.930.521 bp position in chromosome 10, which is in agreement with the BGI result. The control (IV.1-080002) does not have the mutation as expected.

For the "medium" quality variants (c.C2264T and c.C871T) it can be seen the alternate allele in the control (IV.1-080002), which was not reported in the BGI analysis (Figures 10 and 11).

In Figure 10, it is highlighted that the patient III.2-080004 is heterozygous having the reference (G) and alternate (A) alleles for the 13.399.786 bp position in chromosome 3 corresponding to the result reported by BGI. However, in the control (IV.1-080002) the variant was present, although it was supposed to be absent.

In Figure 11, one can see that the patient IV.9-090044 is heterozygous having the reference allele (G) and alternate (A) alleles for the 151.970.931 bp position in chromosome 7 corresponding to the result reported by BGI. However, in the control (IV.1-080002) the variant was supposed to be absent and again it was identified in Sanger sequencing. More, when this position was observed in IGV (Figure 11A) the variant is present in the control, re-confirming these findings and supporting that the results given by the BGI are wrong. Additionally, the strand bias is higher in the "medium" and "bad" quality variants than in the "good" quality variants.

In Table 12 is summarized the inconsistencies between the BGI BAM files, BGI bioinformatics analysis and Sanger sequencing.

Table 12. Differences inspected between IGV, BGI bioinformatics analysis and Sanger sequencing in the QC variants. Here is shown the genotype present for both affected and unnafected individuals when we visualized the BGI BAM files in IGV, the BGI bioinformatics analysis and Sanger sequencing. Discordant results are highlited in bold.

CDS change	Family relative ID	BGI bioinformatics analysis	IGV	Sanger sequencing
c.C8800T	III.6-080005	GA	GA	GA
	IV.1-080002	GG	GG	GG
c.C2264T	III.2-080004	GA	GA	GA
	IV.1-080002	GG	GG	**GA**
c.C871T	IV.9-090044	GA	GA	GA
	IV.1-080002	GG	**GA**	**GA**
c.C478G	III.2-080004	GC	GC	GC
	IV.1-080002	GG	**GC**	NA

CDS change: Coding DNA sequence change; NA: Not available

As the BGI bioinformatics results were analyzed, were found several errors such as old and different versions of dbSNP between files for the same individual, a high rate of non-calling variants and errors in calculations such as the SIFT values. In addition, BGI used GATK Unified Genotyper that has a higher error rate when compared with GATK Haplotype Caller. The latter does a *de novo* realignment of the read data decreasing the error rate of the variants call. Consequently, we decided to perform a new updated bioinformatics analysis based on the Best Practices workflow, from the Broad Institute, optimized for human data, using only the raw FASTQ files from WES sent by BGI.

4.3 New bioinformatics analysis of raw WES data

In the first phase of the bioinformatics re-analysis (Data Pre-processing – Figure 4), FastQC and Qualimap tools were used to perform a quality control of the twelve FASTQ files (six FASTQ files from the positive strand and another six FASTQ files from the negative strand) and the six BAM files. The FastQC and Qualimap were used to annotate the number of reads for each of the six individuals, the GC percentage, mapping and duplication rate (Table 13 and Figure G.1, Appendix G). In addition, was used the graphs sent by BGI to assess the sequencing depth of the target bases and see if there was no variation between samples (Figure G.2, Appendix G).

Table 13. WES data and alignment statistics. This information was generated with the FastQC and Qualimap programmes. The GC, mapping rate and duplicate rate were selected as the most important parameters since they can affect the subsequent analysis.

Data and alignment statistics	III.4-080001	IV.1-080002	III.2-080004	III.6-080005	IV.9-090044	IV.3-090095
Number of reads	63.713.598	58.536.898	59.194.990	50.515.208	58.252.244	75.206.380
Data size	6.371.359.800	5.853.689.800	5.919.499.000	5.051.520.800	5.825.224.400	7.520.638.000
fq1 GC(%)	47.00	47.00	47.00	48.00	46.00	45.00
fq2 GC(%)	47.00	47.00	47.00	48.00	46.00	45.00
Mapping rate (%)	97.83	99.23	98.81	99.21	99.10	98.60
Duplicate rate (%)	22.92	30.90	24.45	27.00	22.51	18.19

fq1 – FASTQ 1 file (positive strand); fq2 – FASTQ 2 file (negative strand)

In the second phase of the bioinformatics analysis (Variant Discovery) to improve the variant calling results, the Ti/Tv ratio was calculated for each of the four variant calling tools (Table 14). Reference values in the literature for Ti/Tv ratio vary between ~2.0 – 2.1 for WGS and ~3.0 – 3.3 for WES[75]. Also, a combined analysis was performed in which all variants that do not match between the four tools were filtered out. The combined analysis achieved a higher Ti/Tv score (2.30) than any of the variant calling tools individually suggesting a lower false positive rate with this analysis. Therefore, the results obtained with the combined analysis were used in the subsequent in SNPs/InDels analyses.

Table 14. Ti/Tv scores. Ti/Tv is a ratio between the number of transitions (A – G, C – T) and the number of transversions (A – C, G – T, A – T, C – G), and is used to assess the false positive rate. This score was calculated for each of the four variant calling tools as well as for a combined analysis. The scores obtained for each tool is an average of the ratios calculated for each of the six individuals sequenced by WES.

Variant Calling tool	Ti/Tv score
UnifiedGenotyper GATK	1.91
HaplotypeCaller GATK	2.19
Freebayes	2.00
Samtools mpileup	1.96
Combined analysis	2.30

Additionally, it was assessed if the QC filters applied to obtain the final lists of SNPs and InDels present in each individual presented optimal results. For this we compared our SNP data with respective data collected for European Americans in the Bamshad *et al.* report[56] (Table 15). The mean number of novel SNPs present in our dataset was 259, approximately the same number obtained in the dataset for the European Americans (novel SNPs = 307). Also, it was calculated the Ti/Tv score per individual to assess the false positive SNP rate (Table 16). A global Ti/Tv score of approximately 2.37 (average score, Table 16) was achieved. This indicates an improved mean error rate when compared to previous calculations and closer to the reference value for WES (3.0). Also, the Ti/Tv score for novel SNPs showed a high error rate when compared to the global Ti/Tv score indicating that these variants could be increasing the global false positive rate.

Table 15. Mean number of novel and known SNPs obtained in the WES re-analysis. Known variants are those found in dbSNP and 1000 GP.

Variant type	Mean number of variants (± SD) in our dataset	Mean number of variants (± SD) in European Americans[56]
Novel variants		
NSV	150 (± 13)	192 (± 21)
Stop-gain and stop-loss	3 (± 0.6)	5 (± 2)
Synonymous	89 (± 10)	109 (± 16)
Splice	17 (± 3)	2 (± 1)
Total	259 (± 16)	307 (± 33)
Known variants		
NSV	8782 (± 160)	9319 (± 233)
Stop-gain and stop-loss	68 (± 4)	89 (± 6)
Synonymous	9968 (± 59)	10.536 (± 280)
Splice	1685 (± 27)	32 (± 3)
Total	20579 (± 219)	19976 (± 505)
Total variants		
NSV	8967 (± 118)	9511 (± 244)
Stop-gain and stop-loss	72 (± 4)	93 (± 6)
Synonymous	10098 (± 103)	10645 (± 286)
Splice	1702 (± 27)	34 (± 4)
Total	20838 (± 219)	20283 (± 523)

SD: Standard Deviation; NSV: Non-synonymous Variants

Note: For the European Americans the dbSNP version used was dbSNP131. For our dataset the dbSNP version used was dbSNP141.

Table 16. Ti/Tv ratio obtained for the six family members sequenced by WES. This score was calculated for each of the six individual sequenced by WES and a mean Ti/Tv score of all six individuals.

	III.4-080001	IV.1-080002	III.2-080004	III.6-080005	IV.9-090044	IV.3-090095	Average Ti/Tv
Ti/Tv	2.37	2.37	2.38	2.43	2.36	2.31	2.37
Novel Ti/Tv	2.11	2.00	2.01	2.29	2.01	2.25	2.11

4.4 Validation by Sanger sequencing

After filtering out the low quality variants and performing the annotation (final steps of the bioinformatics pipeline [Figure 4 – Methods]), six VCF files for the SNP variants and another six VCF files for InDel variants present in the six individuals were generated. Since variants in the CDS region are those of interest in this project, the NSVs and also stop-gain/stop-loss variants for the SNP variants category have been analyzed (Table 17, C.1 and C.3, Appendix C). For InDels category, frameshift insertion/deletion, non-frameshift insertion/deletion, frameshift/non-frameshift block substitution and stop-gain/stop-loss variants have been analyzed (Table 17, C.2 and C.4, Appendix C). The results for each individual were divided by SNPs and InDels in Table 17.

For both SNPs and InDels, the search has been expanded for other regulatory regions (3'UTR, 5'UTR, upstream, downstream, intronic, intergenic and splicing regions – Table 17, C.1, C.2, C.3 and C.4, Appendix C) that which were not validated due to time constrains.

After all the QC and filtering analysis, variants that are absent from the unaffected relative and present in all affected individuals have been identified, so mutations in the same genes shared only among affected family members can be highlight, obtaining a final list of variants to further pursue in validation. However, we did not have a final diagnosis until the end of this thesis of one family member from family 1 (individual IV.3-090095), so two different analyses were created:

- Model 1: one unaffected relative (IV.1-080002) and five affected individuals (III.2-080004, III.4-080001, III.6-080005, IV.3-090095 and IV.9-090044);

- Model 2: two unaffected relatives (IV.1-080002 and IV.3-090095) and four affected individuals (III.2-080004, III.4-080001, III.6-080005 and IV.9-090044).

Assuming that the mutation causing PDB in family 1 is novel, it is not included in the dbSNP or 1000 GP databases. Therefore, all polymorphisms/mutations present in 1000 GP and dbSNP141 were filtered out from our list of variants.

Table 17. SNP and InDel statistics. After our bioinformatics analysis the results obtained for each individual were divided into SNPs (top) and InDels (bottom) variants.

	SNPs					
	III.4-080001	**IV.1-080002**	**III.2-080004**	**III.6-080005**	**IV.9-090044**	**IV.3-090095**
Total variants	107487	105285	105058	91948	109182	163163
1000 GP and dbSNP	101969	100504	99697	88545	103326	153613
1000GP	101969	100504	99697	88545	103326	153613
dbSNP	106433	104210	104053	91038	107998	161669
Novel	1054	1075	1005	910	1184	1494
Hom	55452	51727	51663	40023	53940	94250
Het	52035	53558	53395	51925	55242	68913
Synonymous	10368	10251	10355	10541	10461	10416
NSV	9079	8972	9090	9314	9027	9180
Stop-gain	58	61	57	58	66	54
Stop-loss	12	13	13	13	11	14
Exonic	19518	19298	19516	19927	19565	19665
Exonic and splicing	50	50	52	43	47	61
5'UTR	1869	1788	1779	1891	1764	1862
3'UTR	3603	3537	3530	3616	3692	4433
Upstream	4438	3940	4081	3017	4158	7561
Downstream	2995	2791	2746	1904	2848	5823
Intronic	62028	61231	61154	55448	64701	93275
Intergenic	7863	7794	7258	1897	7452	23280
SIFT	1433	1417	1443	1444	1406	1452
	InDels					
Total variants	7467	8770	7723	6981	7841	9579
1000 GP and dbSNP	4174	4656	3412	3963	4380	5381
1000GP	4174	4656	3412	3963	4380	5381
dbSNP	7212	8451	7468	6761	7619	9307
Novel	255	319	255	220	222	272
Hom	4210	5119	4347	3742	4472	5610
Het	3257	3651	3376	3239	3369	3969
Frameshift insertion	78	72	76	73	70	74
Non-frameshift insertion	98	90	97	106	83	94
Frameshift deletion	71	70	72	69	60	72
Non-frameshift deletion	118	105	119	115	117	102
Frameshift block substitution	0	0	0	0	0	0
Non-frameshift block substitution	0	0	0	0	0	0
Stop-gain	0	0	0	0	0	0
Stop-loss	3	3	2	2	2	3
Exonic	385	357	383	381	349	361
Exonic and splicing	16	20	17	18	14	13
5'UTR	209	156	172	181	159	174
3'UTR	378	347	360	372	388	431
Upstream	295	390	318	246	278	413
Downstream	189	268	185	134	204	305
Intronic	5435	5984	5493	5267	5687	6832
Intergenic	339	989	548	174	518	770

Hom: Homozygous; Het: Heterozygous; NSV: Non-synonymous Variants; SIFT: Sorting Tolerant From Intolerant; Ti/Tv: Transition/Transversion ratio; UTR: Untranslated region.

According to model 1, three novel NSVs, c.C4786T in *KIAA1875*, c.C53T in *NLRC3* and c.T566C in *SRL* co-segregated with the disease in family 1 (Table 18). SIFT and PolyPhen-2 scores indicate that variants c.C4786T and c.T566C are probably damaging for the protein function. In addition, GERP++ scores showed that these two variants are inserted in conserved sites, which means that they are likely to be involved in key biological processes, not tolerating mutations. Only variant c.C53T is predicted to be not damaging for the protein function and also it is inserted in a less conserved site indicating that will tolerate mutations. However, these *in silico* tools are only predictive, so all three variants were validated. The MAPQ for these variants achieved the highest score (MAPQ ~60) indicating a good confidence in the alignment performed. As for coverage, these variants have enough coverage (DP > ~30), however for the c.T566C there is a slight strand bias between the reference and alternate alleles observed in the positive and negative strand (Table 19).

For model 2 one novel NSVs - c.G180A in *SERINC2* - and one novel non-frameshift deletion - c.2163_2168del in *PLEKHG5* - co-segregated with the disease in family 1 (Table 20). SIFT and PolyPhen-2 scores showed ambiguous results for variant c.G180A. The former indicated that the mutation is deleterious and the latter that is benign for the protein function. GERP++ scores highlighted that this variant is inserted in a conserved site, which means that this is inserted in a site that does not tolerate mutations. For InDels there are no *in silico* tools to predict the impact of the mutation. The MAPQ for the two variants achieved the highest score (MAPQ ~60) indicating a good confidence in our alignment. The c.2163_2168del had sufficient coverage (DP ~54) however for c.G180A the coverage is lower (DP ~13). Additionally, there is a slight strand bias between the reference and alternate alleles observed in the positive and negative strand for c.2163_2168del in individuals III.4-080001, III.6-080005 and IV.9-090044 (Table 21).

All mutations for both models were found in the heterozygous state consistent with an autosomal dominant model, since the phenotype is expressed with presence of both reference and alternate alleles.

Table 18. Model 1 top three variants, selected from SNPs and InDels categories from the WES data. These variants are absent from the unaffected relative (IV.1-080002) and present in all five affected individuals (III.2-080004, III.4-080001, III.6-080005, IV.3-090095 and IV.9-090044), and not described in 1000 GP and dbSNP databases.

CDS change	Gene	Mutation type	SIFT score	PolyPhen-2 score	GERP++	Chr	Start/End (bp)	Ref allele	Obs allele	DP	MAPQ	Genotype
c.C4786T	*KIAA1875*	NSV	0.000	0.996	4.46	8	145.171.113/ 145.171.113	C	T	27.00	60.00	0/1
c.C53T	*NLRC3*	NSV	0.920	0.000	-3.62	16	3.627.162/ 3.627.162	G	A	31.00	60.35	0/1
c.T566C	*SRL*	NSV	0.000	0.948	5.10	16	4.245.598/ 4.245.598	A	G	113.00	59.84	0/1

NSV: Non-synonymous Variants; CDS change: Coding DNA sequence change; Chr: Chromosome; bp: Base pair; Ref: Reference; Obs: Observed; SIFT: Sorting Tolerant From Intolerant; Genotype: 1/1 (homozygous for the observed allele); 0/1 (heterozygous).

Table 19. Strand bias for the variants selected according to model 1. The number of reference and alternate alleles seen in IGV in the forward and reverse strand as well as the calculated Fisher strand bias are shown for each model 1 variant and for each individual sequenced by NGS.

		III.4-080001			IV.1-080002			III.2-080004			III.6-080005			IV.9-090044			IV.3-090095		
CDS change	**Strand**	**Ref**	**Obs**	**FS**	**Ref**	**Obs**	**FS**	**Ref**	**Obs**	**FS**	**Ref**	**Obs**	**FS**	**Ref**	**Obs**	**FS**	**Ref**	**Obs**	**FS**
c.C4786T	**Positive**	18	20	1.55	.	.	.	11	12	0.00	13	21	1.62	10	19	0.00	15	14	0.00
	Negative	5	3		.	.		2	2		2	6		2	4		1	2	
c.C53T	**Positive**	17	13	0.00	.	.	.	13	12	1.85	13	14	0.00	12	8	0.00	5	5	0.00
	Negative	9	6		.	.		2	4		5	5		6	3		3	3	
c.T566C	**Positive**	29	38	3.57	.	.	.	38	36	0.73	47	32	7.88	34	43	4.81	38	40	2.35
	Negative	22	21		.	.		19	20		30	11		22	18		20	26	

CDS change: Coding DNA sequence change; Ref: Reference; Obs: Observed; FS: Fisher Strand.

Table 20. Model 2 top three variants, selected from SNPs and InDels categories from the WES data. These variants are absent from the two unaffected relatives (IV.1-080002 and IV.3-090095) and present in four affected individuals (III.2-080004, III.4-080001, III.6-080005 and IV.9-090044), and not described in 1000 GP and dbSNP databases.

CDS change	Gene	Mutation type	SIFT score	PolyPhen-2 score	GERP++	Chr	Start/End (bp)	Ref allele	Obs allele	DP	MAPQ	Genotype
c.G180A	*SERINC2*	NSV	0.030	0.068	4.92	1	31.896.668/ 31.896.668	G	A	13.00	60.00	0/1
c.2163_2168del	*PLEKHG5*	NFD	.	.	.	1	6.529.182/ 6.529.188	TTCCTCC	T	54.00	59.96	0/1

NSV: Non-synonymous Variants; CDS change: Coding DNA sequence change; Chr: Chromosome; bp: Base pair; Ref: Reference; Obs: Observed; SIFT: Sorting Tolerant From Intolerant; Genotype: 1/1 (homozygous for the observed allele); 0/1 (heterozygous).

Table 21. Strand bias for the variants selected according to model 2. The number of reference and alternate alleles seen in IGV in the forward and reverse strand as well as the calculated Fisher strand bias are shown for each model 2 variant and for each individual sequenced by NGS.

		III.4-080001			IV.1-080002			III.2-080004			III.6-080005			IV.9-090044			IV.3-090095		
CDS change	**Strand**	**Ref**	**Obs**	**FS**	**Ref**	**Obs**	**FS**	**Ref**	**Obs**	**FS**	**Ref**	**Obs**	**FS**	**Ref**	**Obs**	**FS**	**Ref**	**Obs**	**FS**
c.G180A	**Positive**	5	4	0.00	.	.	.	2	2	2.76	5	8	1.53	6	6	0.00	.	.	.
	Negative	7	6		.	.		7	2		7	7		5	6		.	.	
c.2163_2168del	**Positive**	10	12	7.86	.	.	.	15	9	0.00	13	8	2.50	11	13	5.91	.	.	.
	Negative	23	11		.	.		14	10		21	9		16	9		.	.	

CDS change: Coding DNA sequence change; Ref: Reference; Obs: Observed; FS: Fisher Strand.

For each variant it was confirmed if the region of interest was amplified - Figures 13, 14, 15, 16 and 17 for both models. Then, the variants for both models were observed in all family members through IGV (Figure 18A, 19A, 20A, 21A and 22A). This way it has been confirmed if the variants were really present in the BAM files (originated after the alignment process) and check the coverage for each mutation.

Afterwards, the validation of these candidate variants were performed by Sanger sequencing and tested the segregation in the remaining family 1 relatives and in family 2. All chromatograms for both models are present in Figures 18B, 19B, 20B, 21B, 22B, 23, 24, 25, 26 and 27 for one strand (complementary strand depicted in Appendix H).

For c.C4786T, one can see that all six patients (family 1) are heterozygous having the reference allele (C) and alternate (T) alleles for the 145.171.113 bp position in chromosome 8 supporting the results obtained by WES analysis (Figure 18B). However, in the control IV.1-080002 the variant was absent in the WES results which is not confirmed as we could see the alteration. In addition, this variant is present in control III.8-090043.

For c.C53T, one can see that all six patients (family 1) are again heterozygous having the reference allele (G) and alternate (A) alleles for the 3.627.162 bp position in chromosome 16 confirming our results (Figure 19B). However, in the control IV.1-080002 the variant was supposed absent in our results but that is not confirmed here as we can see the alteration. In addition, this variant is present in control III.8-090043.

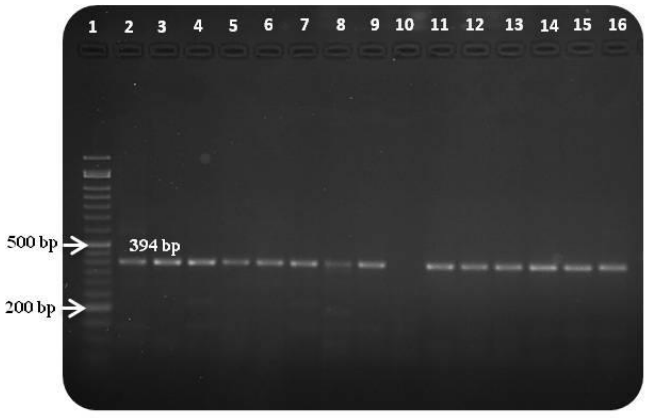
For c.T566C, we can see that all six patients (family 1) are once again heterozygous having the reference allele (A) and alternate (G) alleles for the 4.245.598 bp position in chromosome 16 again confirming the results obtained by us (Figure 20B). However, in the control IV.1-080002 the variant was supposed to be absent according to the bioinformatics analysis, but that was not confirmed by Sanger sequencing. In addition, this variant is present in control III.8-090043.

For c.G180A, we can see that five patients (family 1) are heterozygous as well, having the reference allele (G) and the alternate (A) alleles for the 31.896.668 bp position in chromosome 1 supporting our results (Figure 21B). In the control IV.3-090095 the mutation was not present as it was expected. However, in the control individual IV.1-080002 the variant was supposed to be absent, but that was not confirmed by our sequencing. In addition, this variant is present in control III.8-090043.

For c.2163_2168del we observed that was amplify three instead of one region of interest in the four affected relatives (family 1 - III.1-080003, III.2-080004, III.4-

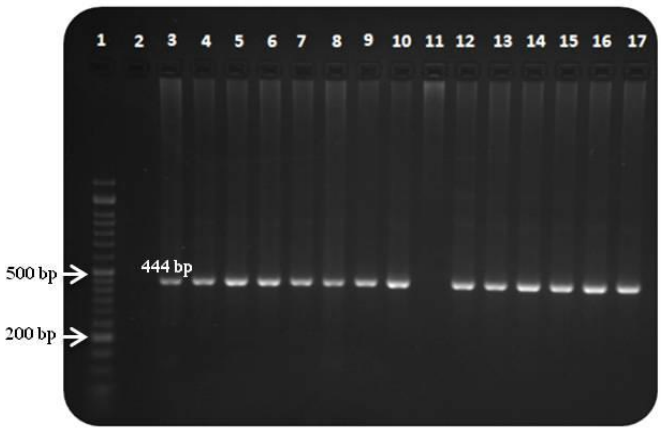
080001, III.6-080005 and IV.9-090044) and also in the control (IV.1-080002, Figure 23). In the case of family 2, more than one region was also amplified, however looking at the agarose gel picture (Figure 17) the distance between the PCR fragments is smaller when compared to the PCR fragments in family 1. Still, this variant was validated, since this contains our region of interest and we want understand what other regions are being amplified. This is discussed in more detail in sub-chapter 5.1.4 One can see in the chromatograms obtained (Figure 22B) that all four patients (family 1) have the deletion and also the control III.8-090043. However, the deletion reported by the WES analysis (TCCTCC – 6.529.182/6.529.188 bp) does not match with the deletion found by Sanger sequencing (CTCCTC – 6.529.184/6.529.190 bp). Additionally, in the control IV.3-090095 the mutation was absent as it was expected. In the control individual IV.1-080002 it was supposed not find the deletion, but that was not confirmed by our sequencing.

Even though the five candidate mutations under study did not segregate as expected with PDB affection status in the Sanger sequencing validation, the presence of these mutations was tested in family 2 to assess their pathogenic potential by replication in an independent family. None of these mutations was detected in any of the family 2 members (Figures 23, 24, 25, 26 and 27). However, a deletion for the 6.529.187/6.529.190 bp position (CTC) in chromosome 1 was observed in the controls (IV.1-140032 and II.1-140033), absent in the patients (II.2-140030 and II.3-140029) and in the unclear individuals (III.1-140031 and III.2-140034).



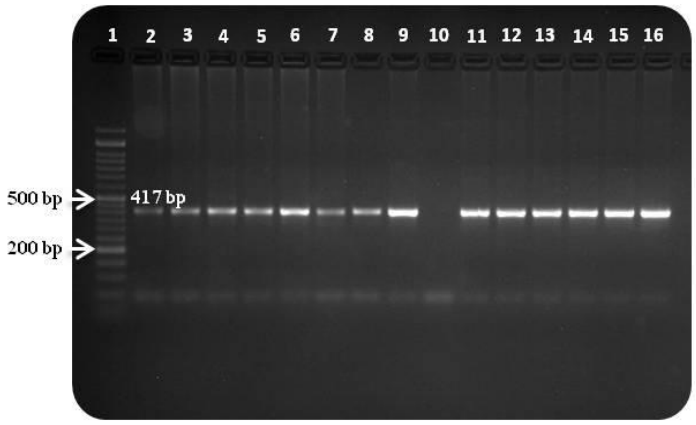
1 – NZYDNALadder VI
2 – III.4-080001
3 – IV.1-080002
4 – III.1-080003
5 – III.2-080004
6 – III.6-080005
7 – IV.3-090095
8 – III.8-090043
9 – IV.9-090044
10 – Negative Control
11 – II.3-140029
12 – II.2-140030
13 – III.1-140031
14 – IV.1-140032
15 – II.1-140033
16 – III.2-140034

Figure 13. PCR amplification of the c.C4786T variant (model 1). This product was run in a 2% agarose gel and is running at the expected location for 349 bp fragments.



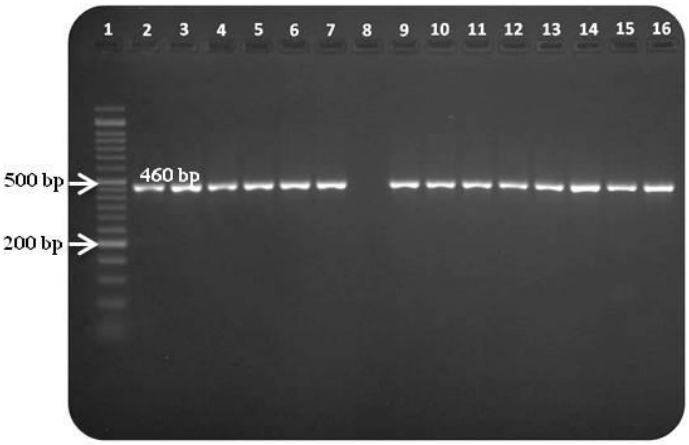
1 – NZYDNALadder VI
2 and 11 – Negative Controls
3 – III.4-080001
4 – IV.1-080002
5 – III.1-080003
6 – III.2-080004
7 – III.6-080005
8 – IV.3-090095
9 – III.8-090043
10 – IV.9-090044
12 – II.3-140029
13 – II.2-140030
14 – III.1-140031
15 – IV.1-140032
16 – II.1-140033
17 – III.2-140034

Figure 14. PCR amplification of the c.C53T variant (model 1). This PCR was run in a 2% agarose gel and is running at the expected location for 444 bp fragment.



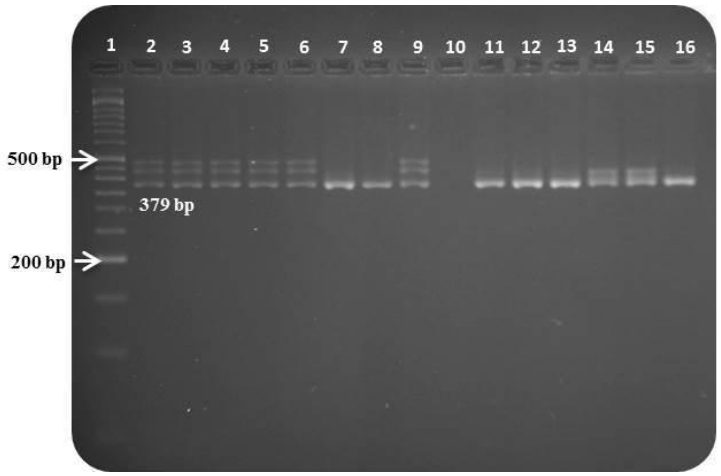
1 – NZYDNALadder VI
2 – III.4-080001
3 – IV.1-080002
4 – III.1-080003
5 – III.2-080004
6 – III.6-080005
7 – IV.3-090095
8 – III.8-090043
9 – IV.9-090044
10 – Negative Control
11 – II.3-140029
12 – II.2-140030
13 – III.1-140031
14 – IV.1-140032
15 – II.1-140033
16 – III.2-140034

Figure 15. PCR amplification of the c.T566C variant (model 1). This product was run in a 2% agarose gel and is running at the expected location for 417 bp fragment.



1 – NZYDNALadder VI
2 – II.3-140029
3 – II.2-140030
4 – III.1-140031
5 – IV.1-140032
6 – II.1-140033
7 – III.2-140034
8 – Negative Control
9 – III.4-080001
10 – IV.1-080002
11 – III.1-080003
12 – III.2-080004
13 – III.6-080005
14 – IV.3-090095
15 – III.8-090043
16 – IV.9-090044

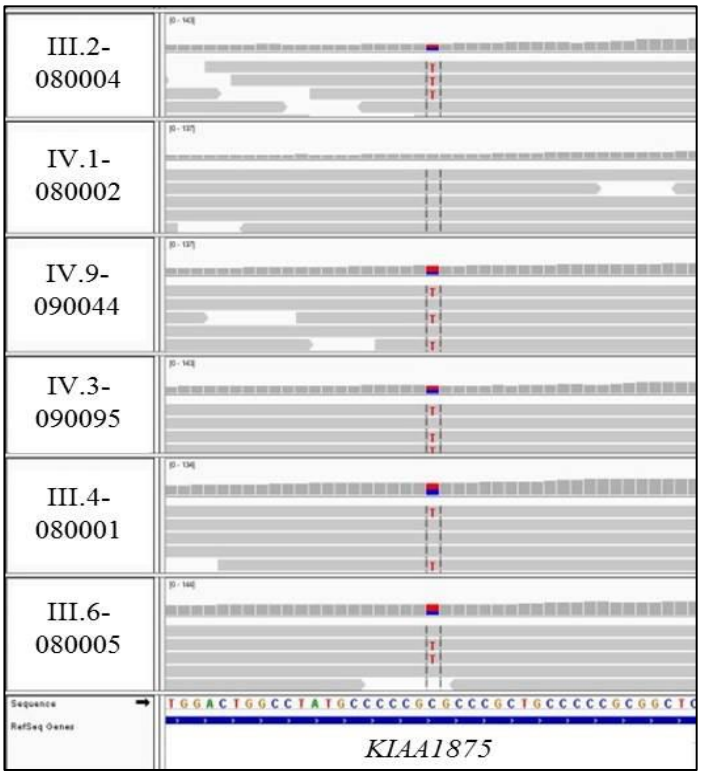
Figure 16. PCR amplification of the c.G180A variant (model 2). This product was run in a 2% agarose gel and is running at the expected location for 460 bp fragment.



1 – NZYDNALadder VI
2 – III.4-080001
3 – IV.1-080002
4 – III.1-080003
5 – III.2-080004
6 – III.6-080005
7 – IV.3-090095
8 – III.8-090043
9 – IV.9-090044
10 – Negative Control
11 – II.3-140029
12 – II.2-140030
13 – III.1-140031
14 – IV.1-140032
15 – II.1-140033
16 – III.2-140034

Figure 17. PCR amplification of the c.2163_2168del variant (model 2). This product was run in a 5% agarose gel and is running at the expected location for 379 bp fragment.

A



T G G C C T A T G C C C C C G [C/T] G C C C C G C T G C C C C C G – Family 1

B

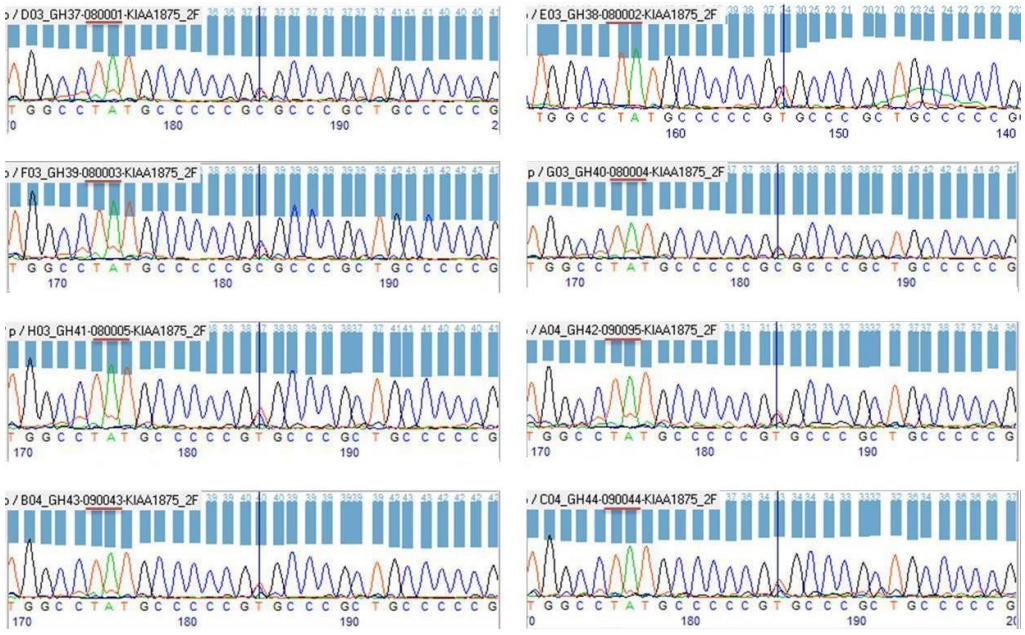
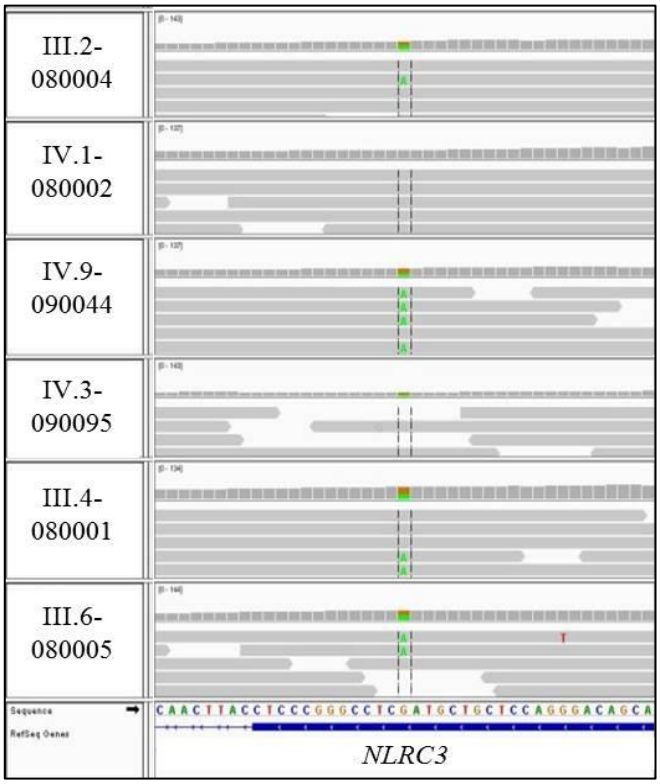


Figure 18. c.C4786T variant according to IGV (A) and chromatograms obtained through Sanger sequencing (forward strand - B). A – Here is shown the mutated allele present in five patients (III.2-080004, III.4-080001, III.6-080005, IV.3-090095 and IV.9-090044) and absent in control (IV.1-080002), via IGV. **B –** Here one can see the mutated T allele for the c.C4786T variant in patients (III.1-080003, III.2-080004, III.4-080001, III.6-080005, IV.3-090095 and IV.9-090044), also present in the controls (IV.1-080002 and III.8-090043). The complementary strands chromatograms are depicted in Appendix H.

A



A C C T C C C G G G C C T C [G /A] A T G C T G G C T C C A G G – Family 1

B

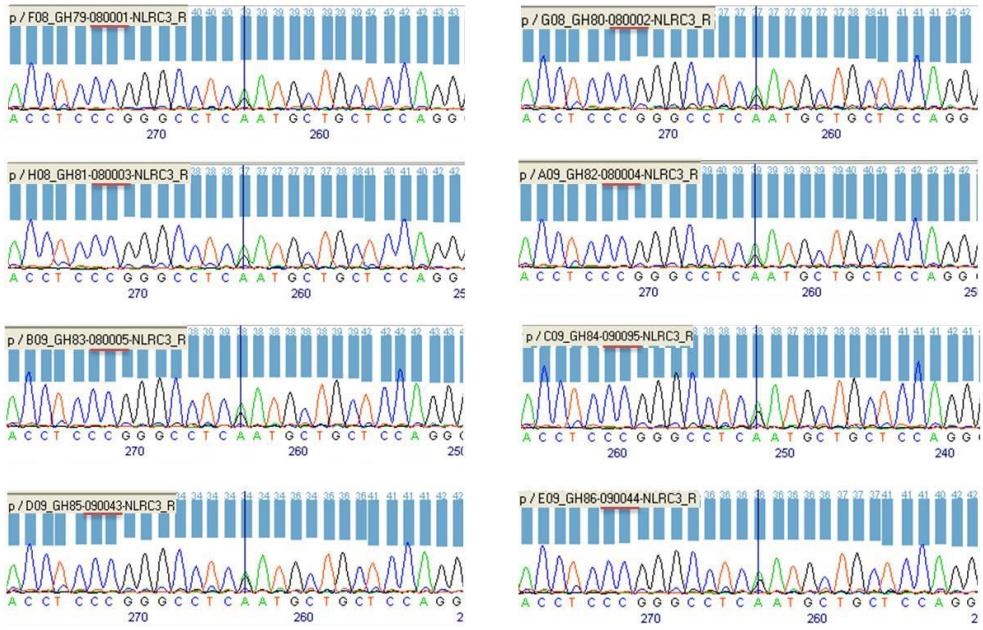
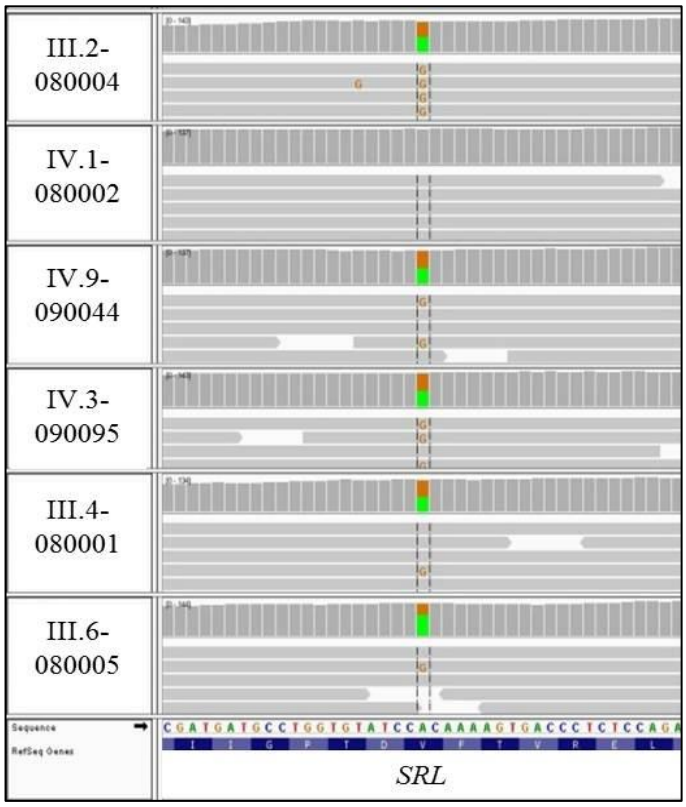


Figure 19. c.C53T variant according to IGV (A) and chromatograms obtained through Sanger sequencing (reverse strand - B). A – Here one can see the mutated allele present in five patients (III.2-080004, III.4-080001, III.6-080005, IV.3-090095 and IV.9-090044) and absent in control (IV.1-080002), through IGV. **B –** Here is shown the mutated A allele for the c.C53T variant in patients (III.1-080003, III.2-080004, III.4-080001, III.6-080005, IV.3-090095 and IV.9-090044), also present in the controls (IV.1-080002 and III.8-090043). The complementary strands chromatograms are presented in the Appendix H.

A



A T G C C T G G T G T A T C C [A/G] C A A A A G T G A C C C T C – Family 1

B

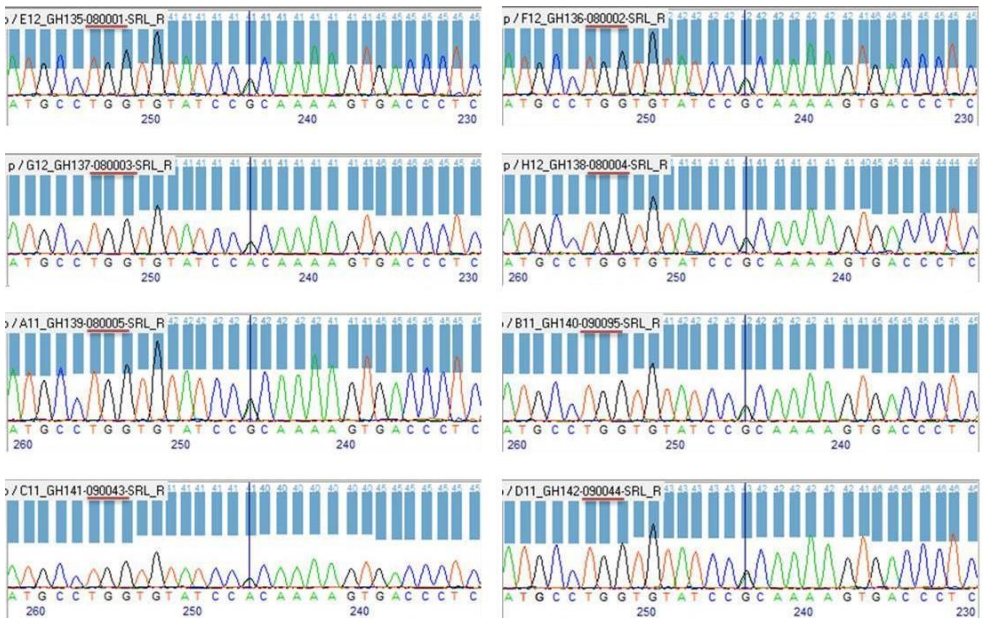
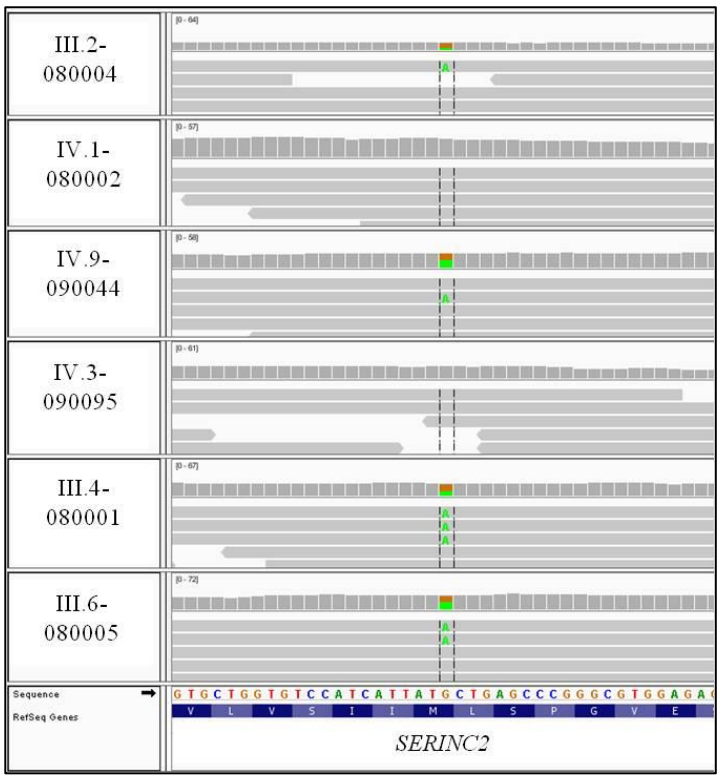


Figure 20. c.T566C variant according to IGV (A) and chromatograms obtained through Sanger sequencing (reverse strand - B). A – Here is shown the mutated allele present in five patients (III.2-080004, III.4-080001, III.6-080005, IV.3-090095 and IV.9-090044) and absent in control (IV.1-080002), through IGV. **B –** Here is shown the mutated G allele for the c.T566C variant in patients (III.1-080003, III.2-080004, III.4-080001, III.6-080005, IV.3-090095 and IV.9-090044), also present in the controls (IV.1-080002 and III.8-090043). The complementary strands chromatograms are presented in the Appendix H.

A



G G T G T C C A T C A T T A T [G/A] C T G A G C C C G G G C G T G – Family 1

B

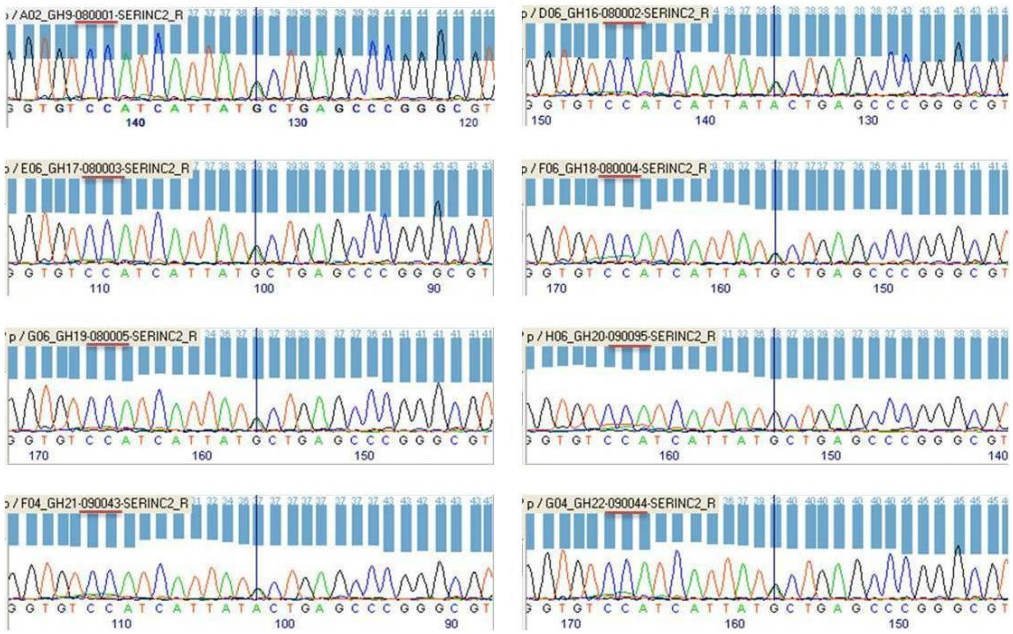
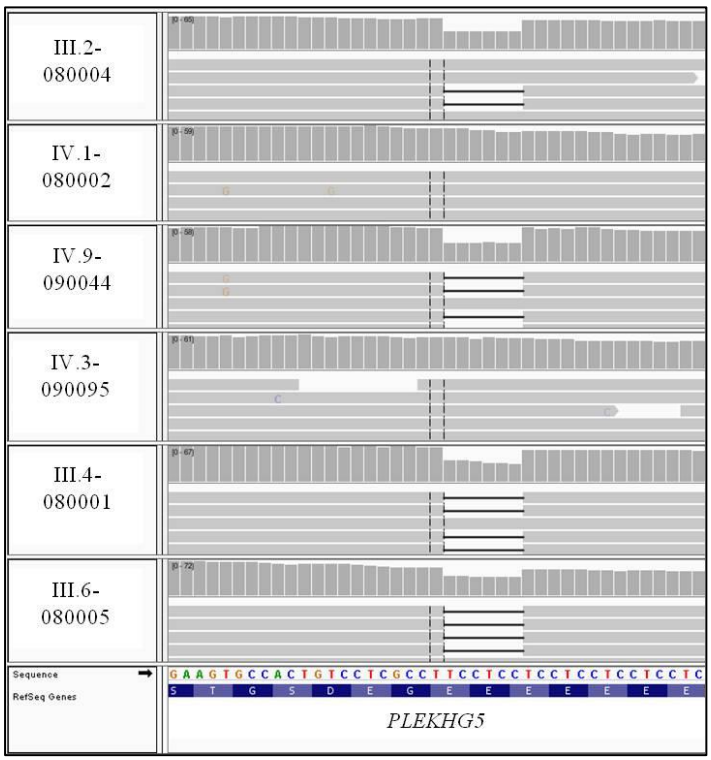


Figure 21. c.G180A variant according to IGV (A) and chromatograms obtained through Sanger sequencing (reverse strand - B). A – Here one can see the mutated allele present in four patients (III.2-080004, III.4-080001, III.6-080005, and IV.9-090044) and absent in controls (IV.1-080002 and IV.3-090095), via IGV. **B –** Here is shown the mutated A allele for the c.G180A variant in patients (III.1-080003, III.2-080004, III.4-080001, III.6-080005 and IV.9-090044), also present in the controls (IV.1-080002 and III.8-090043). The variant is absent in the unclear individual IV.3-090095. The complementary strands chromatograms are presented in the Appendix H.

A



C C T C C T C C T C C T C C T [C C T C C T C / C] T T C C T C C T C C T G C T – Family 1

B

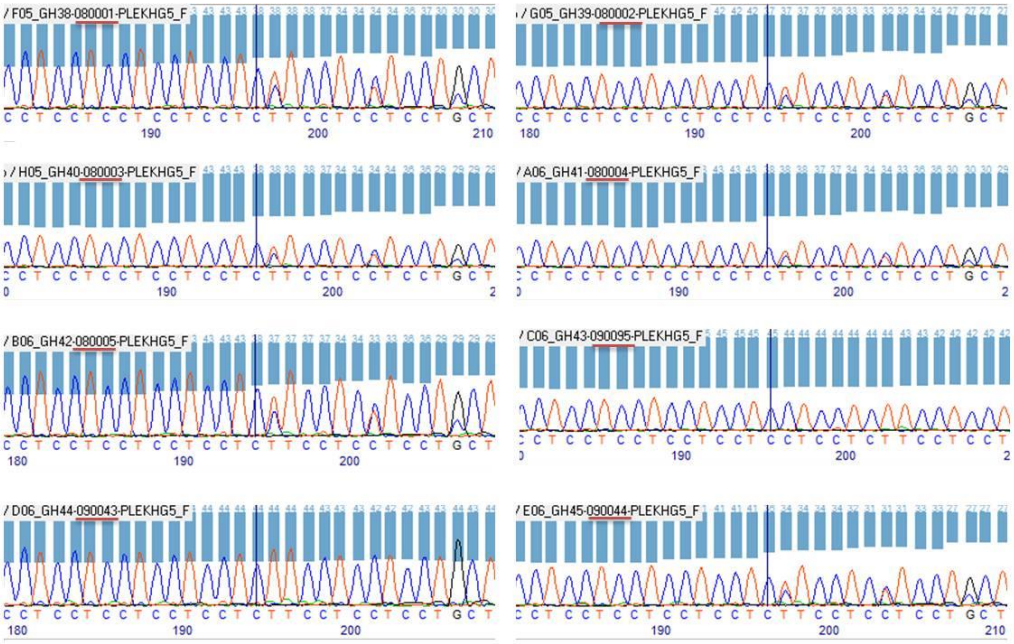


Figure 22. c.2163_2168del variant according to IGV (A) and chromatograms obtained through Sanger sequencing (forward strand - B). A – Here is shown the deletion present in four PDB patients (III.2-080004, III.4-080001, III.6-080005 and IV.9-090044) and absent in two controls (IV.1-080002 and IV.3-090095), through IGV. **B –** Here is shown the deletion CTCCTC for the 6.529.184/6.529.190 bp position in chromosome 1 in patients (III.1-080003, III.2-080004, III.4-080001, III.6-080005 and IV.9-090044), also present in the controls (IV.1-080002 and III.8-090043). The deletion is absent in the unclear individual IV.3-090095. The complementary strands chromatograms are presented in the Appendix H.

T G G C C T A T G C C C C C G [C/T] G C C C C G C T G C C C C C G – Family 2

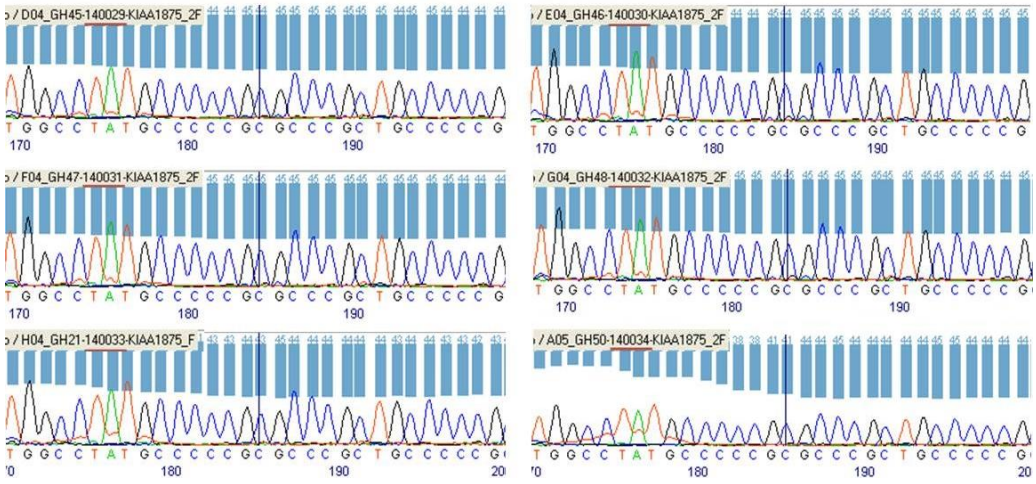


Figure 23. Chromatograms for the c.C4786T variant (forward strand) for family 2. Here one can see the ancestral C allele for the c.C4786T variant in patients (II.2-140030 and II.3-140029), in the unclear individuals (III.1-140031 and III.2-140034) and in the controls (II.1-140033 and IV.1-140032). The complementary strands chromatograms are depicted in Appendix H.

A C C T C C C G G G C C T C [G /A] A T G C T G G C T C C A G G – Family 2

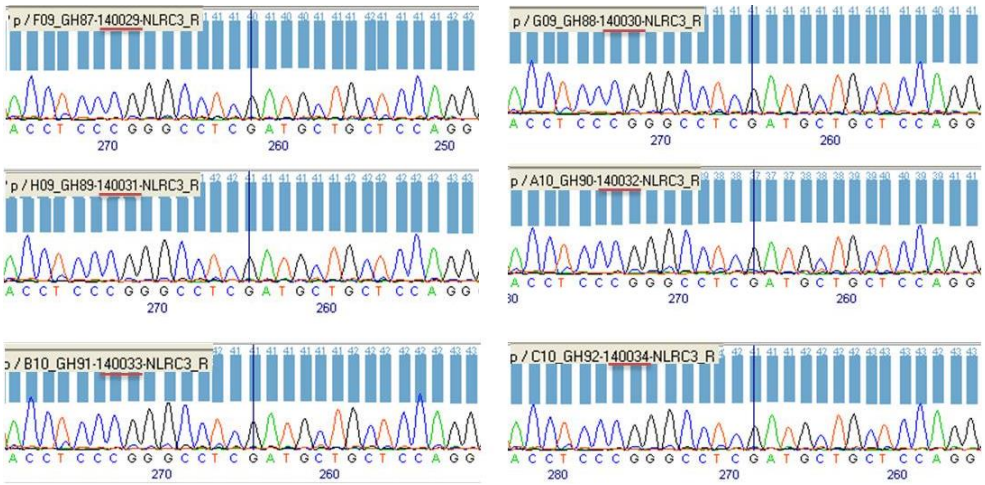


Figure 24. Chromatograms for the c.C53T variant (reverse strand) for family 2. Here is shown the ancestral G allele for the c.C53T variant in patients (II.2-140030 and II.3-140029), in the unclear individuals (III.1-140031 and III.2-140034) and in the controls (II.1-140033 and IV.1-140032). The complementary strands chromatograms are showed in the Appendix H.

A T G C C T G G T G T A T C C [A/G] C A A A A G T G A C C C T C – Family 2

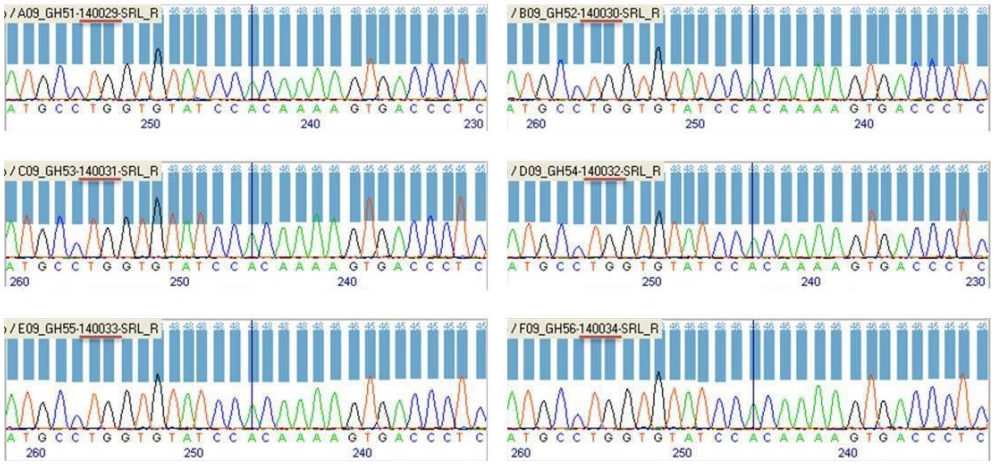


Figure 25. Chromatograms for the c.T566C variant (reverse strand) for family 2. Here is shown the ancestral A allele for the c.T566C variant in patients (II.2-140030 and II.3-140029), in the unclear individuals (III.1-140031 and III.2-140034) and in the controls (II.1-140033 and IV.1-140032). The complementary strands chromatograms are presented in the Appendix H.

G G T G T C C A T C A T T A T [G/A] C T G A G C C C G G G C G T – Family 2

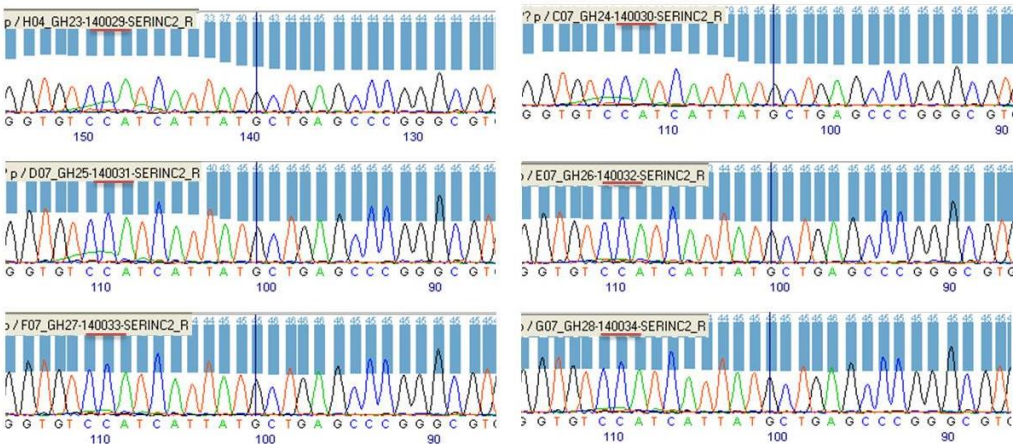


Figure 26. Chromatograms for the c.G180A variant (reverse strand) for family 2. Here one can see the ancestral G allele for the c.G180A variant in patients (II.2-140030 and II.3-140029), in the unclear individuals (III.1-140031 and III.2-140034) and in the controls (II.1-140033 and IV.1-140032). The complementary strands chromatograms are presented in the Appendix H.

C C T C C T C C T C C T C C T [C C T C / C] T T C C T C C T C C T G C T – **Family 2**

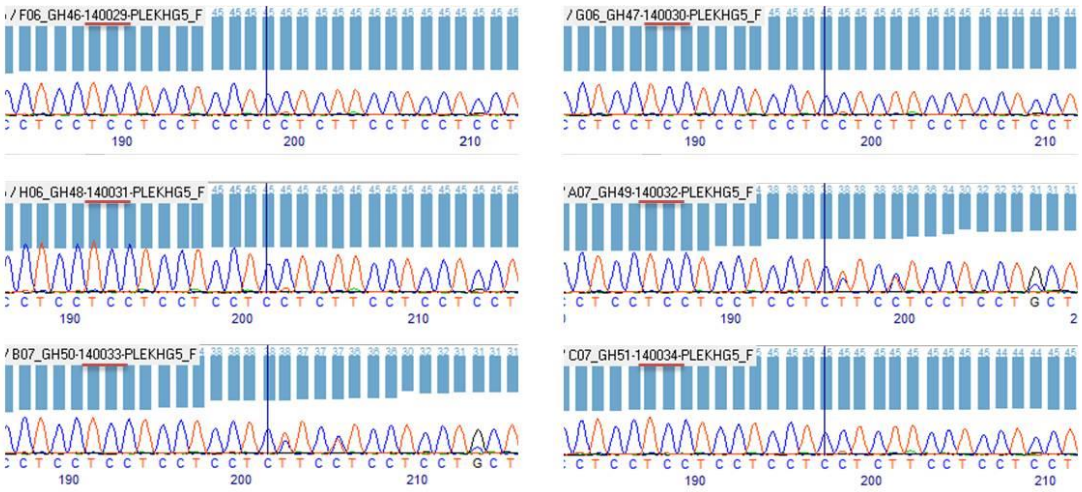


Figure 27. Chromatograms for the c.2163_2168del variant (forward strand) for family 2. Here one can see the deletion CTC for the 6.529.187/6.529.190 bp position in chromosome 1 in the controls (II.1-140033 and IV.1-140032), absent in patients (II.2-140030 and II.3-140029) and in the unclear individuals (III.1-140031 and III.2-140034). The complementary strands chromatograms are showed Appendix H.

In Table 22 is summarized the inconsistencies between the BAM files, bioinformatics analysis and Sanger sequencing.

Table 22. Summary of the differences observed between the WES analysis, bioinformatics re-analysis and the Sanger sequencing. Here is shown the genotype present for both affected and unnafected individuals when we visualized the BAM files in IGV, the bioinformatics analysis and Sanger sequencing. Discordant results are highlited in bold.

CDS change	Family 1 relative ID	Bioinformatics analysis	IGV	Sanger sequencing
c.C4786T	Affected individuals	CT	CT	CT
	Control IV.1-080002	CC	CC	**CT**
c.C53T	Affected individuals	GA	GA	GA
	Control IV.1-080002	GG	GG	**GA**
c.T566C	Affected individuals	AG	AG	AG
	Control IV.1-080002	AA	AA	**AG**
c.G180A	Affected individuals	GA	GA	GA
	Control IV.1-080002	GG	GG	**GA**
	Control IV.3-090095	GG	GG	GG
c.2163_2168del	Affected individuals	TCCTCC/-	TCCTCC/-	**CTCCTC/-**
	Control IV.1-080002	TCCTCC	TCCTCC	**CTCCTC/-**
	Control IV.3-090095	TCCTCC	TCCTCC	TCCTCC

CDS change: Coding DNA sequence change.

4.5 Exonic Variants in PDB-associated genes

Since none of the four novel variants studied previously appeared to co-segregated with PDB with an autosomal dominant mode of inheritance (with complete penetrance), the evolution of variants reported in the dbSNP or 1000 GP databases may have been too restrictive. To test the possibility that the disease-causing mutation in family 1 is in a gene previously associated with PDB (Table 1 and 2) we investigated if any of the variants detected in these genes (Table I.1, Appendix I) co-segregated with PDB in one of the analysis models. Using this selection criteria only one variant in *PML* (c.T1933C) present in an exonic region co-segregated with disease according to model 2, and was further confirmed by Sanger sequencing.

The amplification of the region of interest was confirmed using PCR as before (Figure 28). Then, this variant was observed in all the family members sequenced by WES via IGV (Figure 29). This way it has been confirmed visually if they were really present in the BAM files (originated after the alignment process) and confirm the coverage for each mutation in the six family members. Afterwards, the validation of these variant were performed by Sanger sequencing and tested the segregation in the other relatives from family 1 and in family 2.

The chromatogram is shown in Figure 29B for the reverse strand (and the respective forward strands are in Appendix H). One can see that four patients (family 1) are heterozygous as well, having the reference allele (T) and the alternate (C) alleles for the 74.336.633 bp position in chromosome 15 supporting our results (Figure 29B). In the unclear individual IV.3-090095 the mutation was not identified as it was expected. However, in the control individual IV.1-080002 the variant was supposed to be absent, but that was not confirmed by our sequencing. In addition, the variant was present in control III.8-090043 and in all the relatives (affected and unaffected) from family 2 (Figure 30).

In Table I.1, Appendix I we can see that all six family members sequenced by WES have mutations in genes associated to PDB, even the control (IV.1-080002). Additionally, our bioinformatics pipeline also identified common variants present in intronic, intergenic and splicing regions in our study individuals (results in Table I.1, Appendix I). Since PDB is a complex disease, with incomplete penetrance and several causal genes, it is possible that non affected individuals have the casual genes but do not express the PDB phenotype. Numerous variants are located in regulatory regions. Most

of the NSV were predicted to be not damaging for the protein function the *in silico* SIFT and Polyphen-2 tools (Table I.1, Appendix I).

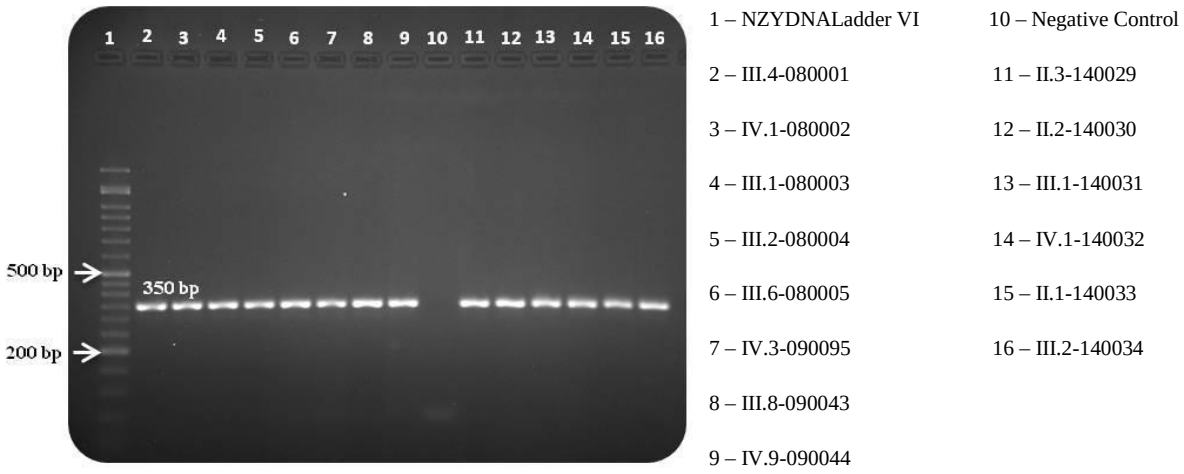
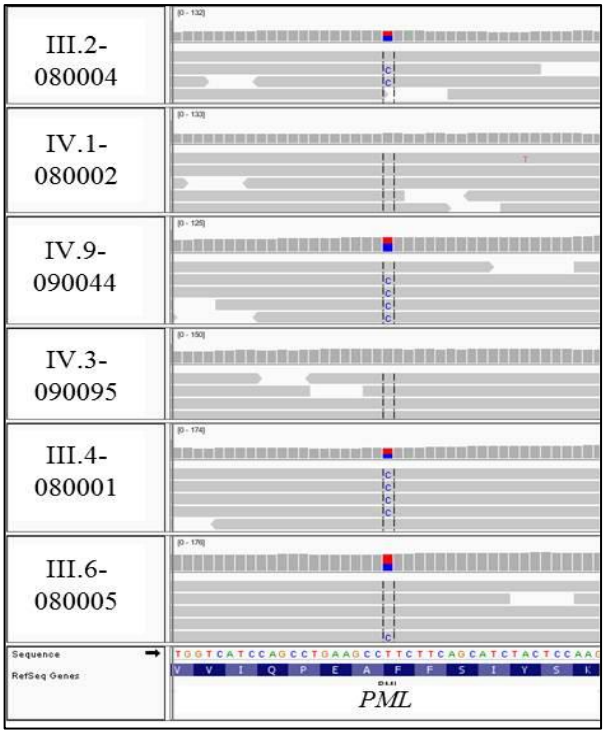


Figure 28. PCR amplification of the c.T1933C variant (model 2). This product was run in a 2% agarose gel and has 350 bp as was expected.

A



A T C C A G C C T G A A G C C [T/C] T C T T C A G C A T C T A C – Family 1

B

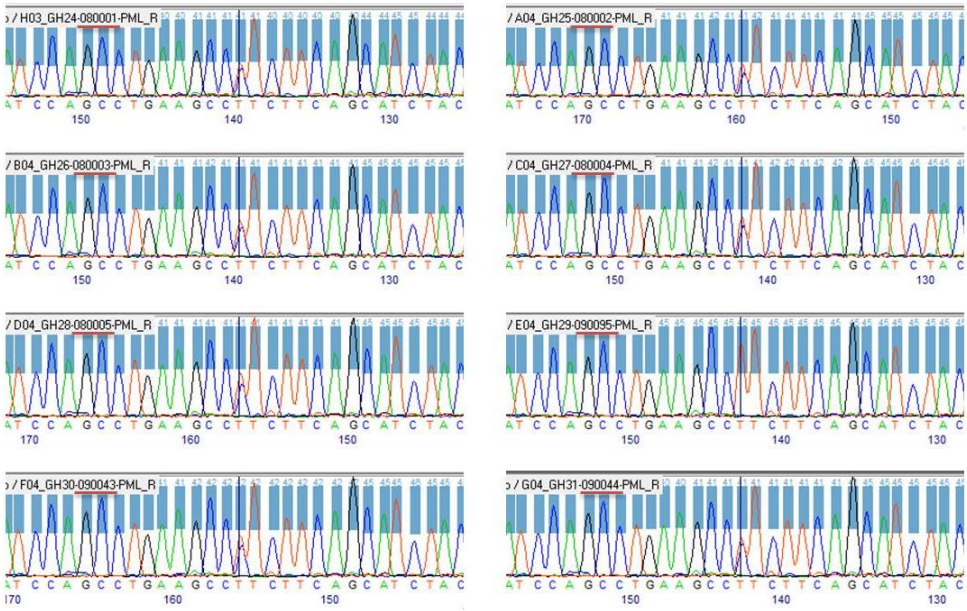


Figure 29. c.T1933C variant according to IGV (A) and chromatograms obtained through Sanger sequencing (reverse strand – B). A - Here one can see the mutated C allele present in the four PDB patients (III.2-080004, III.4-080001, III.6-080005 and IV.9-090044) and absent in the two controls (IV.1-080002 and IV.3-090095), using IGV. **B -** Here is shown the mutated C allele for c.T1933C variant in patients (III.1-080003, III.2-080004, III.4-080001, III.6-080005 and IV.9-090044), also present in controls III.8-090043 and IV.1-080002 from family 1. The variant is absent in the unclear individual IV.3-090095. The complementary strands chromatograms are showed Appendix H.

A T C C A G C C T G A A G C C [T/C] T C T T C A G C A T C T A C – Family 2

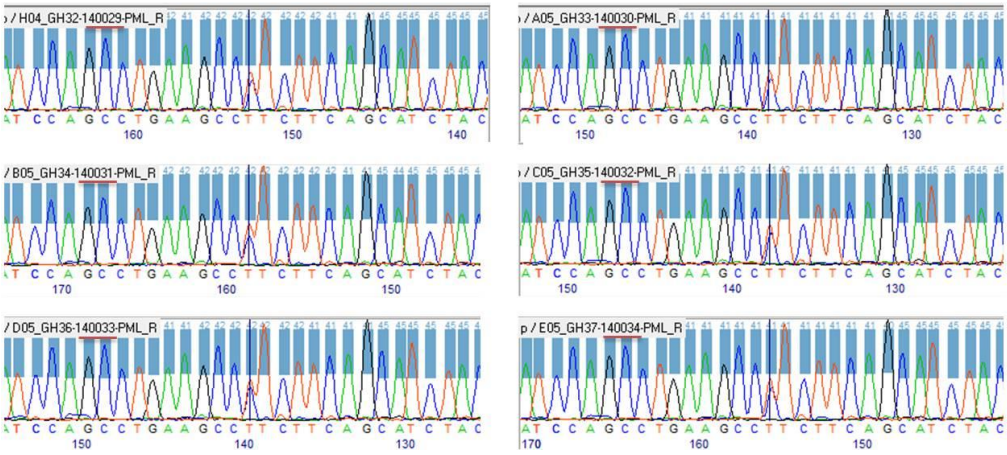


Figure 30. Chromatograms for the c.T1933C variant (reverse strand) for family 2. Here is shown the mutated C allele, for c.T1933C variant in patients (II.2-140030 and II.3-140029), also present in probable PDB affected (III.1-140031 and III.2-140034) and in both controls (II.1-140033 and IV.1-140032) from family 2. The complementary strands chromatograms are showed in the Appendix H.

5. Discussion

5.1 WES data quality control

5.1.1 BGI selected variants

WES is a robust approach to search for disease-related variants, however it still presents with important bioinformatics challenges, such as variant detection and annotation, which ultimately lead to high error rates[61,79]. Moreover, variant annotation tools usually have a strong impact in creating errors due to variant analysis, creating the possibility that we are discarding important variants[97].

Even though chromatograms for the quality control variants (Figures 9B, 10B and 11B) do not have a good quality because the primers used mapped too close to the variant of interest, one can see the presence/absence of the variant. From the four variants selected for the QC, only the "good" quality variant (c.C8800T – Figure 9B) was validated successfully in both individuals. In addition, the result obtained by Sanger sequencing for c.C871T ("medium" quality) for individual IV.1-080002 (control – where the mutation is present – see Figure 11) is confirmed by what one can see using BGI bam files in IGV. Additionally, this variant was present in patient IV9-090044. However, in the VCF files (created by GATK Unified Genotyper) this variant was not present for the control. This contradiction might be explained by the fact that this region has a low MAPQ and did not fulfill the minimum score to proceed in the BGI bioinformatics analysis, most likely being filtered out.

The MAPQ scores for the c.C8800T and c.C2264T variants indicate a good confidence in the alignment (MAPQ ~60), contrasting to c.C871T and c.C478G that have a lower MAPQ (MAPQ ~36.98 and 24.94, respectively). Lastly, c.C478G had a low coverage (DP ~10x), which along with the low MAPQ (~24.94) can be suggestive of a false positive result. One possible explanation for the low MAPQ scores rely on the regions where the variants are inserted being highly repetitive, which can normally result in alignment errors. A possible solution to increase the coverage in regions with low mappability is increasing the read length, thus improving the specificity[98].

There is a balance between the alleles observed in the forward and reverse strands for c.C8800T. However, for c.T1933C, c.C2264T, c.C871T and c.C478G

variants we observed an unbalanced strand bias, which in turn can be an indicator of a false positive SNP. Strand bias can be explained by the fact that c.C2264T is inserted in the extremities of the reads being harder to sequence this region. Also, since the MAPQ for both c.C871T and c.C478G is low, the unbalanced strand bias is possibly due to errors in the read alignment resulting in the observation of more reference or alternate alleles in one of the strands than expected. This way, we can infer that variants with high MAPQ, high coverage, and also with no strand bias are less susceptible to false positive/negative results.

5.1.2 BGI *versus* our bioinformatics pipeline

As carefully analyzed the BGI bioinformatics results several errors were found. Thus, we decided to perform a new entire bioinformatics analysis at this point using only the raw FASTQ files sent by BGI. To minimize errors we performed quality control analyses at three stages of the bioinformatics pipeline: (1) raw data (FASTAQ files), (2) alignment (BAM files) and (3) variant calling (VCF files).

In the first stage (1), we used FastQC to check the GC content and base quality. The GC content for an exome region is usually reported to be between 49 and 51%, and abnormal deviations indicate DNA contamination[79]. We verified that the confidence in our base call is high (Figure G.1, Appendix G) and the GC content is approximately 45 and 48%, indicating that our samples have a good quality to proceed through the pipeline.

In the second stage, we used Qualimap to check both the mapping and duplication rate. These parameters are a good indicator to perceive if the variant call will be accurate. If the mapping and duplication mark is done incorrectly it can affect the subsequent analysis, such as the coverage that could be overestimated if the duplicates do not get marked. The mappability is also a source of false negative variants, being however more challenging to detect. In our dataset approximately 2% of the reads did not map with the human reference (hg19). Since reports indicate that approximately 1-10% of the reads usually do not map with the reference sequence[79], this means that our 2% of unmapped reads are within the expected range. In addition, we used the graphs created by BGI to assess the sequencing depth and we visualized that there is no variation between samples, suggesting that there seems to be no problem with the DNA sequencing itself.

In the third stage, to identify the best approach to proceed in the pipeline, we calculated the Ti/Tv ratio for each of the four variant calling tools used and for the combined analysis. We found that the Unified Genotyper (GATK) and Samtools mpileup had worst scores than the Haplotype Caller (GATK) and Freebayes. Thus, by using the combined analysis we decreased the false positive rate (Ti/Tv ratio = 2.30), since when we combined only the variants that are equal between the four variants calling tools we increased the probability that these variants are true positives. Consequently, we chose the combined analysis to proceed with the bioinformatics analysis.

5.1.3 QC of the variants highlited by our bioinformatics analysis

Efforts to understand association between human genetic variants and their phenotype effects have been made since it is believed that genetic variation is a major factor that stimulates the diversity between individuals. Recent studies have reported that the number of genetic variants in the genome is higher than 3.5 million per individual, making the identification of causative variants an extremely challenging task[36]. Approximately 20.000 SNVs are identified by a WES approach and ~95% of these variants are already known[56].

The number of novel and known SNPs was assessed in the CDS region studied. The number of novel NSVs can be a good indicator of the false positive SNPs present in the study sample[79]. Bamshad *et al.* showed that only 200 novel NSVs are usually expected per individual through a WES approach. Higher numbers are indicative of a high false positive rate. There was ~20.838 SNPs in average called per individual, of which 259 were novel (Table 15). From these, approximately 150 novel SNPs are NSVs against the 192 novel SNPs obtained by Bamshad *et al.* for European Americans (Table 15). Our results indicate that our sample has a low false positive rate as the number of novel NSVs obtained should be smaller since we used a new version of dbSNP (containing a higher number known SNPs).

In addition, we also calculated the Ti/Tv ratio for all the SNPs (known and novel) and also for the novel SNPs only to confirm the rate of false positive in our samples (Table 16). We obtained a Ti/Tv ratio of 2.37 for all SNPs and 2.11 for novel SNPs. When we compared these with the scores from the literature (Ti/Tv of 3.0 for WES[75]) we verify that our rate of false positive is higher. This can be due to the novel

SNPs that reduce the global Ti/Tv ratio obtained for all the variants in the sample. Also, this score can be undervalued because most of the exome capture kits do not capture only the exonic region. So, it is expected the Ti/Tv ratio to vary between 2.0 and 3.0 for SNPs inside these target regions, with the value depending on the fraction of exons inside the target regions[99].

5.1.4 Validation of variants of interest

The most interesting mutations to follow-up (variants present in affected individuals and absent from unaffected relatives), were validated by Sanger sequencing. Introduction of novel next-generation sequencing techniques has dramatically reduced the cost for DNA sequencing, however these techniques have a higher error rate when compared to the tradicional Sanger sequencing[66]. NGS differs from Sanger sequencing in several aspects, including producing 100 times more data in a short time (high-throughput) and depending on the detection of pyrophosphate release on nucleotide incorporation, rather than chain termination with dideoxynucleotides[100]. Furthermore, the NGS technique produces shorter reads (~100bp) but with lower quality when compared to Sanger sequencing[100,101].

MAPQ for the novel variants selected for both models achieved the highest score (MAPQ ~60) indicating good confidence in the alignment. However, the coverage was lower for c.C4786T (DP ~27), c.C53T (DP ~31) and c.G180A (DP ~13) when compared with the coverage achieved for c.T566C variant (DP ~113) and c.2163_2168del (DP ~54). A mean coverage of 60x was requested to BGI to obtain more confident results in the variant calling. A coverage above 30 already indicates a good confidence in the variant call, but below this score the confidence diminishes, suggesting a harder region to sequence. Additionally, for each variant we counted how many reference / observed alleles were seen in the reverse and forward strands to analyze the strand bias (Tables 19 and 21). When the strand bias is significantly different, it is an indication of a false positive call (e.g. for SNPs a strand bias above 60 is excluded from the call set)[79]. A 50% allele balance between the positive (or forward) and negative (or reverse) strands for heterozygous calls is desirable in sequencing data. The bias between the two strands is affected by the preferential reference allele bias caused by the alignment algorithm that penalizes a mismatch from the reference sequence, lowering in turn the percentage of alternate alleles in the reads[79]. For c.T566C

(model 1), we observed a slight strand bias when compared with the other two variants (c.C4786T and c.C53T). One possible explanation to this bias is that novel rare variants are less specific to exome capture itself being more difficult to sequence. Additionally, strand bias can be due to PCR duplications that did not get marked, originating a higher coverage than would be expected. For model 2, we observed a slight strand bias in c.2163_2168del. This region is highly repetitive, which can lead to misalignments.

Our validation stage was very successful in the sense that four variants (c.C4786T, c.C53T, and c.T566C – model 1, and c.G180A – model 2) identified in affected individuals by WES were in fact present and confirmed with our Sanger sequencing. However, it did not validate the WES results for the control individual IV.1-080002 in both models.

In model 1, the regions were succesfully amplified with the variant of interest and the reference and alternate alleles for each variant were observed in all affected individuals in family 1. Nonetheless, these three variants are also present in the control individual IV.1-080002 and the other control individual III.8-090043 (family 1). Moreover, in family 2 these variants are always absent and so are not causal in this family. While looking for novel variants it was already expected that this could happen, since mutations may be different from family to family, particularly because their frequency is unknown in the population. However, since the two families originate from the same Portuguese region (*Alentejo*), it is expected that they may share the same pathogenic mutation through a distant relative (these two families are not known to be related).

For model 2, all PDB patients from family 1 and both controls (IV.1-080002 and III.8-090043) have the c.G180A variant. However, the c.2163_2168del deletion mapped to a different location when compared to its WES results, and is a known deletion already reported in public databases (present both in affected and control individuals according to Sanger sequencing). As for the control IV.3-090095, the variants analyzed were not present (as expected by the WES results), indicating that this individual is not a carrier. Additionally, in family 2, these variants follow the same pattern as in model 1. Still, for model 2 c.2163_2168del, when looked at the agarose gel (Figure 17) it was seen that three instead of one region of interest were amplifying in four affected relatives (family 1 - III.1-080003, III.2-080004, III.4-080001, III.6-080005 and IV.9-090044) and also in the control (IV.1-080002, Figure 17). For family 2, we also observed that more than one region was being amplified but the distance between the

PCR fragments is smaller when compared to the PCR fragments in family 1. However, when we designed the primers we confirmed their uniqueness with BLAT and an *in silico* PCR (UCSC - http://genome.ucsc.edu).

The chromatogram in Figure 22B shows that variant c.2163_2168del is not localized in the amplified region that we were expecting through the bioinformatics analysis (TCCTCC – 6.529.182/6.529.188 bp). Instead, we observed two different deletions in both families. For family 1, the deletion (CTCCTC) was localized in chromosome 1 from 6.529.184 bp until 6.529.190 bp position being present in all affected individuals and also in control IV.1-080002. Additionally, in the control IV.3-090095, we did not identify the mutation as expected. For family 2, we observed a deletion localized from 6.529.187 bp until 6.529.190 bp position (CTC) only present in the healthy controls (II.1-140033 and IV.1-140032). Nonetheless, the three different regions amplified in individuals (family 1 - III.1-080003, III.2-080004, III.4-080001, III.6-080005 and IV.9-090044) and also in the control (IV.1-080002, Figure 23) were not explained with the results obtained. But, for family 2, it is possible that the two regions amplified in individuals IV.1-140032 and II.1-140033 are possibly due to the deletion since the distance between PCR fragments is smaller when compared to the PCR fragments in family 1. Additionaly, we investigated this region in more detail and we found several polymorphisms in the region where the two different deletions are inserted. Two frameshifts and five non-frameshifts are described in this region. Only found reported in dbSNP the deletion present in individuals from family 1 (rs375111412 – c.2143_2148del), but until now there are not reports associating this variant with PDB. InDels are more challenging to detect and validate since the InDel calling is more imprecise and inaccurate[74]. Since c.2143_2148del is too close from the supposed variant found through our bioinformatic analysis (c.2163_2168del), it is possible that the result obtained was due to a misalignment.

False negatives present in control IV.1-080002 have several possible explanations. c.C4786T and c.C53T are located in the extremities of the reads, which result in a lower coverage (< 30). Also, c.G180A has low coverage (DP ~13), which indicates a lower confidence in the variant call. Positions with lower coverage are harder to call with confidence[102]. A technical limitation of the WES approach is that the probes that hybridize with the DNA are designed based on the reference sequence, so they capture preferentially the regions with the reference allele[103]. There are a high number of common SNPs around the variants analyzed that may perturb the sequencing

of these target regions. Regions with a higher number of SNPs when compared to the reference sequence can failed in the hybridization step, consequently being missed in the sequencing[56,103]. Furthermore, besides c.T566C, the rest of the variants analyzed had a high GC content, which prevents proper hybridization and results in a low capture and coverage of these regions[104].

It is also possible that some reads are mapping in the wrong location originating false negative results. This can happen around repeat regions. One approach to avoid these errors is to increase the quality values selected when we use BWA for mapping the reads.

5.2 Variants in genes associated to PDB

There are several genes described in the literature that may play a role in the genetic susceptibility to PDB. We found in all six family relatives analyzed by WES mutations in 11 genes previously associated with PDB (Table I.1, Appendix I). One of the variants that we found was c.T1933C (*PML* - rs5742915). This variant was associated in the literature, and is also present in all four PDB patients and absent from the two unaffected relatives IV.1-080002 and IV.3-090095 through WES analysis. Sanger sequencing confirms the presence of the mutation in all four PDB patients and control individual IV.1-080002 (family 1) and absent in the control IV.3-090095 (Figure 29B). However, for family 2 we observe the variant in all the relatives (affected and unaffected – Figure 30). So, it is possible that *PML* is increasing the risk for PDB in family 1.

The most significant association reported was seen with rs5742915 in *PML* resulting in a phenylalanine to leucine amino acid change at codon 645 of PML[23]. Its function in bone metabolism is at the moment unknown, but it is known to be involved in TGF-β signal, which has a role in the regulation of bone remodeling[23].

Additionally, there are three NSVs in *CSF1*, *NUP205* and *OPTN* present in all family relatives studied, with a MAF < 5%. Since GWAS does not detect variants with MAF < 5% these are not yet reported and would be interesting to validate them in these family relatives and others PDB patients. These three genes are implicated in osteoclast formation (*CSF1*), transporting proteins between the nucleus and cytosol (*NUP205*) and regulation of NFκB signaling and autophagy (*OPTN*)[2,7,24].

Disease severity can be related with the number and type of genetic variants segregating in the family relatives [6]. PDB affected family members can carry different combinations of variants that contribute to disease risk[6]. There are also numerous variants located in regulatory regions within known genes (Table I.1, Appendix I), pointing to a possible relation between these and PDB, so future studies are warranted to clarify their role.

Also, PDB is a complex disease with a late age-at-onset, so healthy controls can have causal variants. This could be affecting our results because the control IV.1-080002 is younger, and so could still develop PDB.

5.3 Novel variants associated to PDB

As previously mentioned, it was necessary create two models to analyze the bioinformatics results since a final diagnosis of one individual (IV.3-090095) was not available until the end of this thesis. Since the interest were in novel rare alleles shared only among affected individuals, variants that are absent from the unaffected relative and present in all affected individuals were selected. Also, variants present in dbSNP and 1000 GP were excluded. Then, the variants putative impact on protein function was assessed using mainly two prediction tools, SIFT and Polyphen-2. Additionally, we watched their level of constraint of our candidate regions to see if they are less conserved, tolerating better mutations when compared to conserved sites, possibly implicating them in key biological processes[84].

For the model 1, we obtained three novel NSVs (Table 18), c.C4786T (*KIAA1875*), c.C53T (*NLRC3*) and c.T566C (*SRL*). Variants c.C4786T and c.T566C were predicted to be probably damaging to the protein function. Moreover, these two variants are inserted in conserved sites, not tolerating mutations and possibly being involved in key biological processes related to PDB. c.C53T was predicted to have no negative impact in protein function and also it is inserted in a less conserved site, which will likely tolerate mutations.

c.4786T is localized at 8q24.3 in *KIAA1875* (exon 24), which has an unknown function. This gene belongs to the WD-repeat domain family which is involved in a wide range of cellular functions, such as gene regulation, vesicular trafficking and cell cycle regulation[105,106]. An association with bone metabolism is yet unknown. The region

amplified (containing 394 bp – Table 3) was investigated and found four SNPs. Two of them (rs4977193 and113230) are present in all family relatives from family 1 and 2. These variants do not change the results obtained in the PCR not being relevant for the study performed.

c.C53T is inserted at 16p13.3 in *NLRC3* (exon 1). The NLRC3 protein is part of the nucleotide-binding leucine-rich repeat (LRR)-containing (NLR) family of sensors, attenuating immune-cell activation by interacting with receptors or their downstream adaptors to inhibit signaling molecules[107,108]. NLRC3 interacts with TRAF6 to attenuate Lys63 (K63)-linked ubiquitination of TRAF6 and activation of NFκB. Since NFκB is expressed in osteoclast progenitor cells and induces osteoclastogenesis by binding to RANKL[109], it is possible that an association between PDB and *NLRC3* could be mediated by an effect in TRAF6, which in turn is an intermediary of the NFκB activation. According to dbSNP[110], for the region amplified (444 bp - Table 4), we found 14 known variants which include one insertion and two deletions. One SNP (rs758747) is present in all family relatives from family 1 and 2. Also, there is one SNP described in databases localized next to our variant. Neverthless, this has no impact in the region amplified and in the results obtained for the variant under study.

c.T566C is also located at 16p13.3, in exon 5 of *SRL*. SRL is located in the lumen of the longintudinal sarcoplasmaticum reticulum, which is associated with the inner side of these membranes through calcium bridges[111]. Reports have shown that Ca^{++}-activated ATPase (correlated with Ca^{++} transport) is part of the Ca^{++}-transport system of sarcoplasmaticum reticulum[112]. A modification in calcium transport could result in changes in bone metabolism, given that it is important for maintaining bone mass. A bone metabolism imbalance could then lead to an increased risk for PDB. According to dbSNP[110] there are 16 SNPs in the region amplified (417 bp – Table 4). SNP (rs10852643) segregates in all relatives from family 2, but in family 1 is absent for individuals III.1-080003, III.2-080004 and III.8-090043. However, this has no impact in the results obtained for the variant under study.

c.C4786T, c.C53T and c.T566C are co-segregating in all affected individuals being also present in unaffected III.8-090043, IV.1-080002 and IV.3-090095. It is possibly that these unaffected individuals are PDB affected but have not yet received a diagnosis or they could be carriers only. These variants can be increasing the risk disease since the genes functions are possibly associated with bone metabolism but only explain a minor portion of the variability.

For model 2, we obtained one novel NSV, c.G180A (*SERINC2*), and one non-frameshift deletion, c.2163_2168del (*PLEKHG5*) (Table 20). For the c.G180A, SIFT and PolyPhen-2 show ambiguous scores. SIFT indicated that the mutation is deleterious and the Polyphen-2 that it is benign for protein function. GERP++ showed that this variant is inserted in a conserved site indicating that it will possibly not tolerate mutations and is probably involved in key biological processes. With the SIFT and GERP++ scores we conclude that probably the mutation is deleterious since it is inserted in a site that do not tolerate mutations.

c.G180A is localized at 1p35.1 in *SERINC2* (exon 24). SERINC2 belongs to the SERINC family of transmembrane proteins that facilitates incorporation of serine into phosphatidylserine and sphingolipids, being related to neural activity and lipid biosynthesis[113,114]. Lipids have an important role in the biomineralization process. The early formation of crystal nuclei within the matrix vesicles is enhanced by the activity of specific enzymes, such as alkaline phosphatase[115]. Moreover, it is reported that phosphatidylserine has an association with calcium deposition and alkaline phosphatase activity. Thus, it is likely that mutations in *SERINC2* can modify the activity of alkaline phosphatase, which is related to PDB causing changes in the biomineralization process leading to deficiencies in bone metabolism. The region amplified (460 bp – Table 4) has 21 known SNPs according to dbSNP[110], absent in all family relatives from family 1 and 2. These polymorphisms have no impact in the results achieved for the variant analyzed.

c.2163_2168del is localized at 1p36.31 in *PLEKHG5* (exon 20). *PLEKHG5* activates the NFκB signaling pathway, which seems to be involved in osteoclastogenesis[116]. Mutations in this gene can lead to modifications in the NFκB pathway, inducing osteoclast activity and resulting in a higher susceptibility to PDB. Several polymorphisms (dbSNP) have been described in the region that we amplified (379 bp – Table 4), which may explain the different results obtained through WES and Sanger sequencing, as discussed in sub-chapter 5.1.4. The 6 bp deletion that are actually present in the affected individuals and control IV.1-080002 from family 1 (c.2143_2148del) is described in dbSNP (rs375111412l), so it does not co-segregate with disease. We also found in the same position a deletion of 3 bp in two healthy controls from family 2 that may be related with a protective function, however more studies are warranted to confirm this.

PDB segregation in family 1 is consistent with an autosomal dominant model of inheritance with incomplete penetrance since all mutations for both models were found in the heterozygous state (with presence of both reference and alternate alleles) and segregate in all family relatives (for family 1), being also present in the two unaffected individuals (III.8-090043 and IV.1-080002). Although we cannot formaly exclude the possibility of an autosomal recessive inheritance of PDB in family 1, it is highly unlikely since there are affected individuals in two consecutive generations, which is more compatible with an autosomal dominant inheritance. Also, the disease segregates in both males and females excluding an X-linked recessive model. This results supports the findings that suggest that familial PDB is inherited in an autosomal dominant fashion[8,15,21,22]. Since we have a family with few individuals genes that cause an autosomal dominant disease are more difficult to identify especially when there is a large number of heterozygous candidate variants[55].

PDB is a complex disease of late onset being more challenging to study since it was difficult to find controls in this family. The family members available in our study only allowed us to select controls who had not reached yet the usual age of disease onset. So, the controls used can still develop the disease in the future. Also, there is a high rate of patients that do not have a diagnosis since they did not develop any symptoms. Moreover, our control IV.1-080002 has not yet reached the age of risk. Thus, we may be assigning a control status to an individual that is actually a PDB patient, and disregarding causal variants in our analysis because of this. Moreover, variants located in regulatory and/or intronic regions can explain part of the genetic variability to PDB and further study on these is now warranted.

Rare variants can have larger effects sizes in complex diseases when compared to common variants identified, helping to identify the causal *loci* [53]. Allele penetrance is also important for the risk of developing the disease, even though it is known that combination of common and rare variants, environmental factors and their interactions are important for the PDB risk[53]. Studies reported that PDB has a highly variable penetrance, so it is possible that unaffected individuals carrying the novel rare-variants identified do not express a phenotype[8,15,21,22]. This may support the unaffected status of individual III.8-090043, since our results shows that she inherited the possible causal alleles and she passed it on to her daughter, which is PDB affected (IV.9-090043).

Also, the causal variant may be outside the exome, not being identified with our WES approach. So, additional studies are needed search for novel rare-variants inserted in regions that were not mapped.

This is the 1st study that aimed to identify novel variants in the exome of PDB families. There is only one study of novel rare genetic variants in PDB to the best of our knowledge, however they only searched for variants located in PDB associated *loci* reported in other studies[25]. We also searched for the Beauregard *et al.* [25] rare variants in our data but did not find any of the variants identified in their PDB study, which is not unexpected since these studies were performed in different populations.

Further studies are needed to explain the possible association between the novel variants identified in this study and PDB, as well as explain their role in PDB pathogenesis.

6. Conclusion/Future work

In this study, we described two Portuguese multiplex PDB families from *Alentejo*. This region as a high prevalence of PDB patients but until now there is no explanation for this fact.

We started to emphasize the importance of the three quality control stages in the bioinformatics analysis of the WES data, since this approach has a high error rate that can impact the power to detect novel rare variants. The error rate can be increased by multiple factors, such as the sample quality, exome capture bias, sequencing errors, mapping errors and PCR duplications not marked. In order to obtain more confident results in the WES approach, the DNA must be freshly extracted prior to analysis. In the bioinformatics analysis, the quality values should be more stringent to avoid detection of false positive/negative variants, especially in the alignment. However, false negatives are more challenging to detect than false positive variants. Other scores should be used, such as VQSLOD (variant quality score recalibration), to assess the WES variants quality and improve the variant call.

We validated the three novel variants identified for model 1 (c.C4786T, c.C53T and c.T566C) and one novel variant for model 2 (c.G180A). For model 1, the three variants are present in all PDB patients from family 1. However, for individual IV.1-080002 and III.8-090043 (control individuals) the variants are also present. Moreover, Sanger sequencing did not confirm the results obtained for individual IV.1-080002 in the WES analysis. For model 2, variant c.2163_2168del (*PLEKHG5*) was possibly misaligned and the real variant observed through Sanger sequencing was in fact a known deletion reported in dbSNP. The variant c.G180A is present in all affected individuals and in both controls (III.8-090043 and IV.1-080002) as before, and it is absent from the additional control (IV.3-090095). Moreover, family 2 present none of the five variants studied. The control IV.1-080002 used to perform this study has not reached the average age of disease onset, so it is possible that she may come to develop the disease or have a misdiagnosis. So, it is possible that we had eliminated variants that can increase the predisposition to PDB. Moreover, control III.8-090043 is possibly a carrier or was misdiagnosed since present all the variants analyzed and also has a daughter with PDB. So, we may be assigning a control status to family relatives that are actually PDB patients.

Taken all togheter, one can conclude the results obtained support the hypothesis that point out for familial PDB to be inherited in an autosomal dominant mode with incomplete penetrance.

Future studies should extend to low-frequency variants (MAF between 1% and 5%) present in dbSNP (which contain a large number of pathogenic alleles). Also, the study sample should be bigger to improve the identification of causal variants and with healthy controls above the average age-at-onset. There are also numerous variants located in regulatory regions pointing to a possible presence of novel variants implicated in this region, so future studies are warranted to clarify their role in PDB.

Further studies are now warranted to discover if the four novel variants identified are in fact absent from further true controls individuals and if so, to uncover their role in PDB pathogenesis.

Bibliography

1. Marieb, E. N., Wilhelm, P. B. & Mallatt, J. *Human Anatomy*. 122 (Pearson Benjamin Cummings, 2010).

2. Chung, P. Y. J. & Van Hul, W. Paget's disease of bone: evidence for complex pathogenetic interactions. *Semin. Arthritis Rheum.* **41,** 619–41 (2012).

3. Good, D. a *et al.* Identification of SQSTM1 mutations in familial Paget's disease in Australian pedigrees. *Bone* **35,** 277–82 (2004).

4. Hocking, L. J. *et al.* Genomewide search in familial Paget disease of bone shows evidence of genetic heterogeneity with candidate loci on chromosomes 2q36, 10p13, and 5q35. *Am. J. Hum. Genet.* **69,** 1055–61 (2001).

5. Sofaer, J. a, Holloway, S. M. & Emery, a E. A family study of Paget's disease of bone. *J. Epidemiol. Community Heal.* **37,** 226–231 (1983).

6. Helfrich, M. H. & Hocking, L. J. Genetics and aetiology of Pagetic disorders of bone. *Arch. Biochem. Biophys.* **473,** 172–82 (2008).

7. Ralston, S. H. & Layfield, R. Pathogenesis of Paget disease of bone. *Calcif. Tissue Int.* **91,** 97–113 (2012).

8. Reddy, S. V. Etiology of Paget's disease and osteoclast abnormalities. *J. Cell. Biochem.* **93,** 688–96 (2004).

9. Ralston, S. H., Langston, A. L. & Reid, I. R. Pathogenesis and management of Paget's disease of bone. *Lancet* **372,** 155–63 (2008).

10. Takata, S. *et al.* Guidelines for diagnosis and management of Paget's disease of bone in Japan. *J. Bone Miner. Metab.* **24,** 359–67 (2006).

11. Selby, P. ., Davie, M. W. ., Ralston, S. . & Stone, M. . Guidelines on the management of Paget's disease of bone*. *Bone* **31,** 366–373 (2002).

12. Chung, P. Y. J. *et al.* Indications for a genetic association of a VCP polymorphism with the pathogenesis of sporadic Paget's disease of bone, but not for TNFSF11 (RANKL) and IL-6 polymorphisms. *Mol. Genet. Metab.* **103,** 287–92 (2011).

13. Gennari, L. *et al.* SQSTM1 gene analysis and gene-environment interaction in Paget's disease of bone. *J. Bone Miner. Res.* **25,** 1375–84 (2010).

14. Siris, E. S., Ottman, R., Flaster, E. & Kelsey, J. L. Familial aggregation of Paget's disease of bone. *J. Bone Miner. Res.* **6,** 495–500 (1991).

15. Jones, J. V. & Mervyn F. Reed. Paget's Disease: a family with six cases. 90–91 (1967).

16. Michou, L., Collet, C., Laplanche, J.-L., Orcel, P. & Cornélis, F. Genetics of Paget's disease of bone. *Joint. Bone. Spine* **73,** 243–8 (2006).

17. Rhodes, E. C. *et al.* Sequestosome 1 (SQSTM1) mutations in Paget's disease of bone from the United States. *Calcif. Tissue Int.* **82,** 271–7 (2008).

18. Bolland, M. J. & Cundy, T. Paget's disease of bone: clinical review and update. *Postgrad. Med. J.* **90,** 328–31 (2014).

19. Laurin, N. *et al.* Paget disease of bone: mapping of two loci at 5q35-qter and 5q31. *Am. J. Hum. Genet.* **69,** 528–43 (2001).

20. Rendina, D. *et al.* Evidence for Increased Clinical Severity of Familial and Sporadic Paget ' s Disease of Bone in Campania , Southern Italy. **21,** 1828–1835 (2006).

21. Hiruma, Y. *et al.* A SQSTM1/p62 mutation linked to Paget's disease increases the osteoclastogenic potential of the bone microenvironment. *Hum. Mol. Genet.* **17,** 3708–19 (2008).

22. Hocking, L. *et al.* Familial Paget's disease of bone: patterns of inheritance and frequency of linkage to chromosome 18q. *Bone* **26,** 577–580 (2000).

23. Albagha, O. M. E. *et al.* Genome-wide association identifies three new susceptibility loci for Paget's disease of bone. *Nat. Genet.* **43,** 685–9 (2011).

24. Albagha, O. M. E. *et al.* Genome wide association study identifies variants at CSF1 , OPTN and TNFRSF11A as genetic risk factors for Paget ' s disease of bone. *Eur. PMC Funders Gr.* **42,** 520–524 (2010).

25. Beauregard, M. *et al.* Identification of rare genetic variants in novel loci associated with Paget's disease of bone. *Hum. Genet.* **133,** 755–68 (2014).

26. Tilyardt, M. W. *et al.* A Probable Linkage Between Familial Paget ' s Disease and the HLA Loci *. 498–500 (1982).

27. Cody, J. D. *et al.* Genetic linkage of Paget disease of the bone to chromosome 18q. *Am. J. Hum. Genet.* **61,** 1117–22 (1997).

28. Good, D. a *et al.* Linkage of Paget disease of bone to a novel region on human chromosome 18q23. *Am. J. Hum. Genet.* **70,** 517–25 (2002).

29. Kovach, M. J. *et al.* Clinical delineation and localization to chromosome 9p13.3-p12 of a unique dominant disorder in four families: hereditary inclusion body myopathy, Paget disease of bone, and frontotemporal dementia. *Mol. Genet. Metab.* **74,** 458–75 (2001).

30. Donáth, J. *et al.* Vitamin D receptor, oestrogen receptor-alpha and calcium-sensing receptor genotypes, bone mineral density and biochemical markers in Paget's disease of bone. *Rheumatology (Oxford).* **43,** 692–5 (2004).

31. Wuyts, W. *et al.* Evaluation of the Role of RANK and OPG Genes in Paget ' s Disease of Bone. **28,** 104–107 (2001).

32. Chamoux, E. *et al.* The p62 P392L mutation linked to Paget's disease induces activation of human osteoclasts. *Mol. Endocrinol.* **23,** 1668–80 (2009).

33. Kurihara, N. *et al.* Contributions of the Measles Virus Nucleocapsid Gene and the SQSTM1/p62^(P392L) Mutation to Paget's Disease. *Natl. Institutes Heal.* **13,** 23–34 (2012).

34. Michou, L. *et al.* Gene expression profile in osteoclasts from patients with Paget's disease of bone. *Bone* **46,** 598–603 (2010).

35. Mullaney, J. M., Mills, R. E., Pittard, W. S. & Devine, S. E. Small insertions and deletions (INDELs) in human genomes. *Hum. Mol. Genet.* **19,** R131–6 (2010).

36. Wu, J. & Jiang, R. Prediction of deleterious nonsynonymous single-nucleotide polymorphism for human diseases. *ScientificWorldJournal.* **2013,** 675851 (2013).

37. Carlson, C. S., Eberle, M. a, Kruglyak, L. & Nickerson, D. a. Mapping complex disease loci in whole-genome association studies. *Nature* **429,** 446–52 (2004).

38. Teare, M. D., Barrett, J. H., Road, B. H. & Sheffield, S. Genetic Epidemiology 2 - Genetic linkage studies. (2005).

39. Greenberg, D. a, Abreu, P. & Hodge, S. E. The power to detect linkage in complex disease by means of simple LOD-score analyses. *Am. J. Hum. Genet.* **63,** 870–9 (1998).

40. Ralston, S. H. Pathogenesis of Paget's disease of bone. *Bone* **43,** 819–25 (2008).

41. Daroszewska, A. & Ralston, S. H. Genetics of Paget's disease of bone. *Clin. Sci. (Lond).* **109,** 257–63 (2005).

42. Hughes, a E. *et al.* Mutations in TNFRSF11A, affecting the signal peptide of RANK, cause familial expansile osteolysis. *Nat. Genet.* **24,** 45–8 (2000).

43. Haslam, S. I. *et al.* Paget ' s Disease of Bone : Evidence for a Susceptibility Locus on Chromosome 18q and for Genetic Heterogeneity. **13,** (1998).

44. Michou, L. *et al.* Novel SQSTM1 mutations in patients with Paget's disease of bone in an unrelated multiethnic American population. *Bone* **48,** 456–60 (2011).

45. Risch, N. & Merikangas, K. The future of genetic studies of complex human diseases. *Science* **273,** 1516–7 (1996).

46. Lee, S., Abecasis, G. R., Boehnke, M. & Lin, X. Rare-Variant Association Analysis: Study Designs and Statistical Tests. *Am. J. Hum. Genet.* **95,** 5–23 (2014).

47. Rao, D. C. An overview of the genetic dissection of complex traits. *Adv. Genet.* **60,** 3–34 (2008).

48. Chen, W.-M. & Abecasis, G. R. Family-based association tests for genomewide association scans. *Am. J. Hum. Genet.* **81,** 913–26 (2007).

49. Lewallen, S. & Courtright, P. Epidemiology in practice: case-control studies. *Community Eye Health* **11,** 57–8 (1998).

50. Daroszewska, A. *et al.* Susceptibility to Paget's disease of bone is influenced by a common polymorphic variant of osteoprotegerin. *J. Bone Miner. Res.* **19,** 1506–11 (2004).

51. Beyens, G. *et al.* Identification of Sex-Specific Associations Between Polymorphisms of the OsteoprotegerinGene,TNFRSF11B, and Paget's Disease of Bone. **22,** 1062–1071 (2007).

52. Neale, S. D., Schulze, E., Smith, R. & Athanasou, N. A. The influence of serum cytokines and growth factors on osteoclast formation in Paget ' s disease. 233–240 (2002).

53. Asimit, J. & Zeggini, E. Testing for rare variant associations in complex diseases. *Genome Med.* **1,** 24 (2011).

54. Vasli, N. & Laporte, J. Impacts of massively parallel sequencing for genetic diagnosis of neuromuscular disorders. *Acta Neuropathol.* **125,** 173–85 (2013).

55. Boycott, K. M., Vanstone, M. R., Bulman, D. E. & Mackenzie, A. E. Rare-disease genetics in the era of next-generation sequencing: discovery to translation. *Nat. Rev. Genet.* **14,** 681–91 (2013).

56. Bamshad, M. J. *et al.* Exome sequencing as a tool for Mendelian disease gene discovery. *Nat. Rev. Genet.* **12,** 745–55 (2011).

57. Ng, S. B. *et al.* Exome sequencing identifies the cause of a mendelian disorder. *Nat. Genet.* **42,** 30–5 (2010).

58. Ng, S. B. *et al.* Targeted capture and massively parallel sequencing of 12 human exomes. *Nature* **461,** 272–6 (2009).

59. Hoischen, A. *et al.* De novo mutations of SETBP1 cause Schinzel-Giedion syndrome. *Nat. Genet.* **42,** 483–5 (2010).

60. Jia, P. *et al.* Consensus rules in variant detection from next-generation sequencing data. *PLoS One* **7,** e38470 (2012).

61. Smith, K. R. *et al.* Reducing the exome search space for mendelian diseases using genetic linkage analysis of exome genotypes. *Genome Biol.* **12,** R85 (2011).

62. Haines, J. L. & Pericak-Vance, M. A. *Approaches to Gene Mapping in Complex Human Diseases*. (Wiley-Liss, 1998).

63. Clark, D. P. *Molecular Biology*. chap. 21, 567–598 (Elsevier Academic Press Publications, 2005).

64. Cavaluzzi, M. J. & Borer, P. N. *Revised UV extinction coefficients for nucleoside-5'-monophosphates and unpaired DNA and RNA*. (Nucleic Acids Research 32, 2004).

65. Agilent 2100 Bioanalyzer. in *Agil. Technol.* at <http://www.chem.agilent.com/library/usermanuals/Public/G2946-90004_Vespucci_UG_eBook_(NoSecPack).pdf>

66. Crona, J. *et al.* Next-generation sequencing in the clinical genetic screening of patients with pheochromocytoma and paraganglioma. *Endocr. Connect.* **2,** 104–111 (2013).

67. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25,** 1754–60 (2009).

68. Thorvaldsdóttir, H., Robinson, J. T. & Mesirov, J. P. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Brief. Bioinform.* **14,** 178–92 (2013).

69. Yi, M. *et al.* Performance comparison of SNP detection tools with illumina exome sequencing data-an assessment using both family pedigree information and sample-matched SNP array data. *Nucleic Acids Res.* **42,** e101 (2014).

70. http://www.broadinstitute.org/gatk/events/3391/GATKw1310-BP-5-Variant_calling.pdf.

71. https://wiki.gacrc.uga.edu/wiki/Freebayes.

72. http://samtools.sourceforge.net/mpileup.shtml.

73. McKenna, A. *et al.* The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20,** 1297–303 (2010).

74. O'Rawe, J. *et al.* Low concordance of multiple variant-calling pipelines: practical implications for exome and genome sequencing. *Genome Med.* **5,** 28 (2013).

75. DePristo, M. a *et al.* A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.* **43,** 491–8 (2011).

76. http://www.broadinstitute.org/gatk/gatkdocs/org_broadinstitute_sting_gatk_walkers_variantrecalibration_VariantRecalibrator.html.

77. http://www.broadinstitute.org/gatk/gatkdocs/org_broadinstitute_sting_gatk_walkers_annotator_QualByDepth.html.

78. http://www.broadinstitute.org/gatk/gatkdocs/org_broadinstitute_sting_gatk_walkers_annotator_RMSMappingQuality.html.

79. Guo, Y., Ye, F., Sheng, Q., Clark, T. & Samuels, D. C. Three-stage quality control strategies for DNA re-sequencing data. *Brief. Bioinform.* (2013). doi:10.1093/bib/bbt069

80. http://www.broadinstitute.org/gatk/gatkdocs/org_broadinstitute_sting_gatk_walkers_annotator_Coverage.html.

81. Abecasis, G. R. *et al.* A map of human genome variation from population-scale sequencing. *Nature* **467,** 1061–73 (2010).

82. Kumar, P., Henikoff, S. & Ng, P. C. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nat. Protoc.* **4,** 1073–81 (2009).

83. Adzhubei, I. a *et al.* A method and server for predicting damaging missense mutations. *Nat. Methods* **7,** 248–9 (2010).

84. Hodgkinson, A. *et al.* Selective constraint, background selection, and mutation accumulation variability within and between human populations. *BMC Genomics* **14,** 495 (2013).

85. Davydov, E. V *et al.* Identifying a high fraction of the human genome to be under selective constraint using GERP++. *PLoS Comput. Biol.* **6,** e1001025 (2010).

86. Craigen, W. J. *et al.* Exome sequencing of a patient with suspected mitochondrial disease reveals a likely multigenic etiology. *BMC Med. Genet.* **14,** 83 (2013).

87. Wang, K., Li, M. & Hakonarson, H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res.* **38,** e164 (2010).

88. Ogino, S., Gulley, M. L., den Dunnen, J. T. & Wilson, R. B. Standard mutation nomenclature in molecular diagnostics: practical and educational challenges. *J. Mol. Diagn.* **9,** 1–6 (2007).

89. Untergasser, A. *et al.* Primer3--new capabilities and interfaces. *Nucleic Acids Res.* **40,** e115 (2012).

90. Butler, J. M. *Forensic DNA Typing - Biology Technology and genetics of STR Markers*. (2005).

91. Ishiguro, A. *et al.* High-throughput detection of multiple genetic polymorphisms influencing drug metabolism with mismatch primers in allele-specific polymerase chain reaction. *Anal. Biochem.* **337,** 256–61 (2005).

92. Hecker, K. H. & Roux, K. H. High and low annealing temperatures increase both specificity and yield in touchdown and stepdown PCR. *Biotechniques* **20,** 478–85 (1996).

93. KAPA2G Robust HotStart PCR Kit - Technical Data Sheet. 1–4 (2014).

94. Staden, R. *The Staden sequence analysis package*. 233–41 (Mol Biotechnol 5, 1996).

95. http://gatkforums.broadinstitute.org/discussion/1268/how-should-i-interpret-vcf-files-produced-by-the-gatk.

96. http://www.broadinstitute.org/gatk/gatkdocs/org_broadinstitute_sting_gatk_walkers_annotator_DepthPerAlleleBySample.html.

97. Mccarthy, D. J. *et al.* Choice of transcripts and software has a large effect on variant annotation. (2014).

98. Sims, D., Sudbery, I., Ilott, N. E., Heger, A. & Ponting, C. P. Sequencing depth and coverage: key considerations in genomic analyses. *Nat. Rev. Genet.* **15,** 121–32 (2014).

99. Guo, Y. *et al.* The effect of strand bias in Illumina short-read sequencing data. *BMC Genomics* **13,** 666 (2012).

100. Pareek, C. S., Smoczynski, R. & Tretyn, A. Sequencing technologies and genome sequencing. *J. Appl. Genet.* **52,** 413–35 (2011).

101. Day-Williams, A. G. & Zeggini, E. The effect of next-generation sequencing technology on complex trait research. *Eur. J. Clin. Invest.* **41,** 561–7 (2011).

102. Powers, S., Gopalakrishnan, S. & Tintle, N. Assessing the impact of non-differential genotyping errors on rare variant tests of association. *Hum. Hered.* **72,** 153–60 (2011).

103. Jiang, T., Yang, L., Jiang, H., Tian, G. & Zhang, X. High-performance single-chip exon capture allows accurate whole exome sequencing using the Illumina Genome Analyzer. *Sci. China. Life Sci.* **54,** 945–52 (2011).

104. Sule, G. *et al.* Next-generation sequencing for disorders of low and high bone mineral density. **24,** 2253–2259 (2013).

105. Smith, T. F. Diversity of WD-Repeat Proteins. (2008).

106. Gori, F., Friedman, L. G. & Demay, M. B. Wdr5, a WD-40 protein, regulates osteoblast differentiation during embryonic bone development. *Dev. Biol.* **295,** 498–506 (2006).

107. NLRC3 gene. at <http://omim.org/entry/615648>

108. Schneider, M. *et al.* The innate immune sensor NLRC3 attenuates Toll-like receptor signaling via modification of the signaling adaptor TRAF6 and transcription factor NF-κB. **13,** 823–831 (2013).

109. Ikeda, T., Utsuyama, M. & Hirokawa, K. Expression profiles of receptor activator of nuclear factor kappaB ligand, receptor activator of nuclear factor kappaB, and osteoprotegerin messenger RNA in aged and ovariectomized rat bones. *J. Bone Miner. Res.* **16,** 1416–25 (2001).

110. Sherry, S. T. *et al.* dbSNP : the NCBI database of genetic variation. **29,** 308–311 (2001).

111. SRL gene. at <http://www.omim.org/entry/604992>

112. MacLennan, D. H. & Wong, P. T. Isolation of a calcium-sequestering protein from sarcoplasmic reticulum. *Proc. Natl. Acad. Sci. U. S. A.* **68,** 1231–5 (1971).

113. Inuzuka, M., Hayakawa, M. & Ingi, T. Serinc, an activity-regulated protein family, incorporates serine into membrane lipid synthesis. *J. Biol. Chem.* **280,** 35776–83 (2005).

114. SERINC2 gene. at <http://www.omim.org/entry/614549>

115. Merolli, A. & Santin, M. Role of phosphatidyl-serine in bone repair and its technological exploitation. *Molecules* **14,** 5367–81 (2009).

116. PLEKHG5 gene. at <http://www.omim.org/entry/611101>

Appendices

Appendix A - Scripts used in the bioinformatics analysis

1. Data pre-processing

a) Raw reads (FASTQ files) – FastQC

(http://www.bioinformatics.babraham.ac.uk/projects/download.html#fastqc) to all files

b) Map to reference with BWA

Prepare reference genome file to BWA (approximately 1h20min)

1 bwa index hg19.fa

Align FASTQ1 with reference genome (approximately 44 min)

2 bwa aln -o1 -e50 -m100000 -t4 -i15 -q10 -I hg19.fa fastaq1.fq > fastaq1.fq.aln

Align FASTQ2 with reference genome

3 bwa aln -o1 -e50 -m100000 -t4 -i15 -q10 -I hg19.fa fastaq2.fq > fastaq2.fq.aln

With FASTQ1 and FASTQ2 aligned, combine the two FASTQ files and create a sam file (use the sampe option when reads are pair-end) (approximately 31 min)

4 bwa sampe hg19.fa fastaq1.fq.aln fastaq2.fq.aln fastaq1.fq fastaq2.fq > P1.sam

c) Mark duplicates and sort with Picard

Convert from sam to bam

5 samtools view -S -b P1.sam > P1.bam

6 java -jar picard-tools-1.111/SortSam.jar VALIDATION_STRINGENCY=LENIENT INPUT=P1.bam OUTPUT=P1.sort_picard.bam SORT_ORDER=coordinate

Mark duplicates

7 java -jar picard-tools-1.111/MarkDuplicates.jar INPUT=P1.sort_picard.bam OUTPUT=P1rmdup.bam METRICS_FILE=P1metricsfile.txt VALIDATION_STRINGENCY=LENIENT

d) InDel realignment using GATK (and Samtools & Picard for some steps)

Create a .fai filev for GATK RealignerTargetCreator

8 samtools faidx hg19.fa

Create a .dict filev for GATK RealignerTargetCreator

9 java -jar picard-tools-1.111/CreateSequenceDictionary.jar R=hg19.fa O=hg19.dict

Add ReadGroup

10 java -jar picard-tools-1.111/AddOrReplaceReadGroups.jar INPUT=P1rmdup.bam OUTPUT=P1readgroup.bam PL=illumina LB=unknown PU=unknown SM=unknown VALIDATION_STRINGENCY=LENIENT

11 java -jar picard-tools-1.111/ReorderSam.jar I=P1readgroup.bam O=P1reorder.bam REFERENCE=hg19.fa VALIDATION_STRINGENCY=LENIENT

Create a index file for GATK RealignerTargetCreator

12 java -jar picard-tools-1.111/BuildBamIndex.jar INPUT=P1reorder.bam VALIDATION_STRINGENCY=LENIENT

Download of known sites for indel realignment

ftp://ftp.broadinstitute.org/bundle/2.8/hg19/

GATK RealignerTargetCreator

13 java –jar GATK_2.4.3/GenomeAnalysisTK-2.4-3-g2a7af43/GenomeAnalysisTK.jar -T RealignerTargetCreator -R hg19.fa -I P1reorder.bam -o P1realigner.intervals –known Mills_and_1000G_gold_standard.indels.hg19.vcf

Indel Realigner

14 java -jar GATK2.4.3/GenomeAnalysisTK-2.4-3-g2a7af43/GenomeAnalysisTK.jar -T IndelRealigner -R hg19.fa -I P1reorder.bam –known Mills_and_1000G_gold_standard.indels.hg19.vcf -targetIntervals P1realigner.intervals -o P1realigned.bam

e) Base recalibration using GATK

Builds recalibration model

15 java -jar GATK2.4.3/GenomeAnalysisTK-2.4-3-g2a7af43/GenomeAnalysisTK.jar -T BaseRecalibrator -R hg19.fa -I P1realigned.bam -knownSites dbsnp_138.hg19.chrM_reordered.vcf -knownSites Mills_and_1000G_gold_standard.indels.hg19.vcf -o P1recal.table

Print reads

16 java -jar GATK2.4.3/GenomeAnalysisTK-2.4-3-g2a7af43/GenomeAnalysisTK.jar -T PrintReads -R hg19.fa -I P1realigned.bam -BQSR P1recal.table -o P1recal.bam

Base recalibrator - second pass, which evaluates how the data looks like after recalibration

17 java -jar GATK2.4.3/GenomeAnalysisTK-2.4-3-g2a7af43/GenomeAnalysisTK.jar -T BaseRecalibrator -R hg19.fa -I P1realigned.bam -knownSites dbsnp_138.hg19.chrM_reordered.vcf -knownSites Mills_and_1000G_gold_standard.indels.hg19.vcf -BQSR P1recal.table -o P1after_recal.table

2. Variant discovery

2.1 Unified Genotyper - GATK

2.1.1 SNPs

a) Variant calling

1 java -jar GATK_2.4.3/GenomeAnalysisTK-2.4-3-g2a7af43/GenomeAnalysisTK.jar -T UnifiedGenotyper -R hg19.fa -I P1recal.bam -o P1.vcf -stand_call_conf 50 -stand_emit_conf 10.0 -A Coverage -A RMSMappingQuality -baq CALCULATE_AS_NECESSARY

2.1.2 InDels

a) Variant calling

2 java -jar GATK_2.4.3/GenomeAnalysisTK-2.4-3-g2a7af43/GenomeAnalysisTK.jar -T UnifiedGenotyper -R hg19.fa -I P1recal.bam -o P1indel.vcf -stand_call_conf 50 -stand_emit_conf 10.0 -A Coverage -A RMSMappingQuality -baq CALCULATE_AS_NECESSARY -glm INDEL

The previous step (2) has given an error so the procedure from PrintReads was repeated with the –DIQ option

3 java -jar GATK_2.4.3/GenomeAnalysisTK-2.4-3-g2a7af43/GenomeAnalysisTK.jar -T PrintReads -R hg19.fa -I P1realigned.bam -BQSR P1recal.table -o P1recal_indel.bam –DIQ

VariantCalling - with PrintReads -DIQ

4 java -jar GATK_2.4.3/GenomeAnalysisTK-2.4-3-g2a7af43/GenomeAnalysisTK.jar -T UnifiedGenotyper -R hg19.fa -I P1recal_indel.bam -o P1indel.vcf -stand_call_conf 50 -stand_emit_conf 10.0 -A Coverage -A RMSMappingQuality -baq CALCULATE_AS_NECESSARY -glm INDEL

2.2 Haplotype Caller – GATK

a) Variant calling

5 java -jar GATK_3.1.1/GenomeAnalysisTK.jar -T HaplotypeCaller -R hg19.fa -I P1recal.bam -o P1_HC.vcf -stand_call_conf 50 -stand_emit_conf 10 -minPruning 3 -A Coverage -A RMSMappingQuality

2.2.1 SNPs

a) Select variants

6 java -jar Paget/Softwares/GATK_2.4.3/GenomeAnalysisTK-2.4-3-g2a7af43/GenomeAnalysisTK.jar -T SelectVariants -R hg19.fa -V P1_HC.vcf -selectType SNP -o P1_raw_snps.vcf

2.2.2 InDels

a) Select variants

7 java -jar GATK_2.4.3/GenomeAnalysisTK-2.4-3-g2a7af43/GenomeAnalysisTK.jar -T SelectVariants -R hg19.fa -V P1_HC.vcf -selectType INDEL -o P1_raw_indels.vcf

2.3 Freebayes

a) Variant calling (approximately 10h)

create bam.bai file

8 samtools index P1recal.bam

9 freebayes -f hg19.fa -b P1recal.bam -v P1_freebayes.vcf -0 --genotype-qualities

2.3.1 SNPs

a) Select variants

10 java -jar GATK_2.4.3/GenomeAnalysisTK-2.4-3-g2a7af43/GenomeAnalysisTK.jar -T SelectVariants -R hg19.fa -V P1_freebayes.vcf -selectType SNP -o P1_fb_raw_snps.vcf

2.3.2 InDels

a) Select variants

11 java -jar GATK_2.4.3/GenomeAnalysisTK-2.4-3-g2a7af43/GenomeAnalysisTK.jar -T SelectVariants -R hg19.fa -V P1_freebayes.vcf -selectType INDEL -o P1_fb_raw_indels.vcf

2.4 Samtools mpileup

a) Variant calling (approximately 6h)

12 samtools mpileup -C50 -ugf hg19.fa P1recal.bam | bcftools view -bvcg - > P1_samtools.bcf

Convert from bcf to vcf

13 bcftools view P1_samtools.bcf > P1_samtools.vcf

Select variants - SNPs

14 java -jar GATK_2.4.3/GenomeAnalysisTK-2.4-3-g2a7af43/GenomeAnalysisTK.jar -T SelectVariants -R hg19.fa -V P1_samtools.vcf -selectType SNP -o P1_raw_snps_sam.vcf

Select variants - INDELs

15 java -jar GATK_2.4.3/GenomeAnalysisTK-2.4-3-g2a7af43/GenomeAnalysisTK.jar -T SelectVariants -R hg19.fa -V P1_samtools.vcf -selectType INDEL -o P1_raw_indels_sam.vcf

2.5 Combined analysis

2.5.1 SNPs

Compare the output files obtained with the 4 variant call tools and create a final file only with the common variants between the 4 tools, for each individual – to construct a combined analysis

16 awk 'NR==FNR{a[$2];next} $2 in a‘P1_fb_raw_snps.vcf P1_HCfilter_snp_chrA.vcf > P1_1.vcf

17 awk 'NR==FNR{a[$2];next} $2 in a' P1_raw_snps_sam.vcf P1filter_snp_chrA.vcf > P1_2.vcf

18 awk 'NR==FNR{a[$2];next} $2 in a' P1_1.vcf P1_2.vcf > P1_snps.vcf

b) VariantRecalibrator

19 java -jar GATK_3.1.1/GenomeAnalysisTK.jar -T VariantRecalibrator -R hg19.fa -input P1_snps.vcf -resource:hapmap,known=false,training=true,truth=true,prior=15.0 hapmap_3.3.hg19.chrM_reordered.vcf -resource:omni,knwon=false,training=true,truth=true,prior=12.0 1000G_omni2.5.hg19.vcf -resource:1000G,known=false,training=true,truth=false,prior=10.0 Mills_and_1000G_gold_standard.indels.hg19.vcf -resource:dbsnp,known=true,training=false,truth=false,prior=2.0 dbsnp_138.hg19.chrM_reordered.vcf -an DP -an QD -an FS -an MQ -an MQRankSum -mode SNP -recalFile P1raw.SNPs.recal -tranchesFile P1raw.SNPs.tranches -rscriptFile P1recalSNP.plots.R

ApplyRecalibration

20 java -jar GATK_3.1.1/GenomeAnalysisTK.jar -T ApplyRecalibration -R hg19.fa -input P1_snps.vcf -mode SNP --ts_filter_level 99.0 -recalFile P1raw.SNPs.recal -tranchesFile P1raw.SNPs.tranches -o P1recalibrated_snps.vcf

c) VariantFiltration

21 java -jar GATK_3.1.1/GenomeAnalysisTK.jar -T VariantFiltration -R hg19.fa -V P1recalibrated_snps.vcf --filterExpression "QD < 2.0 || FS > 60.0 || MQ < 40.0 || HaplotypeScore > 13.0 || MQRankSum < -12.5 || ReadPosRankSum < -8.0" --filterName "Hard_to_Validate" -o P1filter_snps_1.vcf

22 java -jar GATK_3.1.1/GenomeAnalysisTK.jar -T VariantFiltration -R hg19.fa -V P1filter_snps_1.vcf --filterExpression "DP < 4" --filterName "Low_Read_Support" -o P1filter_snps_2.vcf

Remove variants present in chrX, chrY and chrM

23 cat P1filter_snps_2.vcf | grep -v '^chrX' | grep -v '^chrY' | grep -v '^chrM' | grep -v 'Hard_to_Validate' | grep -v 'Low_Read_Support' | grep -v 'LowQual' > P1_snp_chrA_combined.vcf

2.5.2 InDels

Compare the output files obtained with the 4 variant call tools and create a final file only with the common variants between the 4 tools, for each individual – to construct a combined analysis

24 awk 'NR==FNR{a[$2];next} $2 in a' P1_fb_raw_indels.vcf P1_HCfilter_indel_chrA.vcf > P1_1.vcf

25 awk 'NR==FNR{a[$2];next} $2 in a' P1_raw_indels_sam.vcf P1filter_indel_chrA.vcf > P1_2.vcf

26 awk 'NR==FNR{a[$2];next} $2 in a' P1_1.vcf P1_2.vcf > P1_indels.vcf

b) VariantRecalibrator

27 java -jar GATK_3.1.1/GenomeAnalysisTK.jar -T VariantRecalibrator -R hg19.fa -input P1_indels.vcf -resource:mills,known=true,training=true,truth=true,prior=12.0 Mills_and_1000G_gold_standard.indels.hg19.vcf -an DP -an FS -an MQRankSum -an ReadPosRankSum -mode INDEL --maxGaussians 4 -recalFile P1recalibrated_INDEL.recal -tranchesFile P1recalibrated_INDEL.tranches -rscriptFile P1recalibrated_INDEL.plots.R

ApplyRecalibration

29 java -jar GATK_3.1.1/GenomeAnalysisTK.jar -T ApplyRecalibration -R hg19.fa -input P1_indels.vcf -mode INDEL --ts_filter_level 99.0 -recalFile P1recalibrated_INDEL.recal -tranchesFile P1recalibrated_INDEL.tranches -o P1recalibrated_indel.vcf

c) VariantFiltration

30 java -jar GATK_3.1.1/GenomeAnalysisTK.jar -T VariantFiltration -R hg19.fa -V P1recalibrated_indel.vcf --filterExpression "QD < 2.0 || FS > 200.0 || ReadPosRankSum < -20.0" --filterName "Hard_to_Validate" -o P1filter_indels_1.vcf

31 java -jar GATK_3.1.1/GenomeAnalysisTK.jar -T VariantFiltration -R hg19.fa -V P1filter_indels_1.vcf --filterExpression "DP < 4" --filterName "Low_Read_Support" -o P1filter_indels_2.vcf

Remove variants present in chrX, chrY and chrM

32 cat P1filter_indels_2.vcf | grep -v '^chrX' | grep -v '^chrY' | grep -v '^chrM' | grep -v 'Hard_to_Validate' | grep -v 'Low_Read_Support' | grep -v 'LowQual' > P1_indel_chrA_combined.vcf

3. Variant annotation

3.1 Functional analysis

Link to download dbNSFP in SnpEff

http://sourceforge.net/projects/snpeff/files/databases/dbNSFP/dbNSFP2.4.txt.gz/download

http://sourceforge.net/projects/snpeff/files/databases/dbNSFP/dbNSFP2.4.txt.gz.tbi/download

Link to dbSNP

ftp://ftp.ncbi.nih.gov/snp/organisms/human_9606_b141_GRCh37p13/VCF/All.vcf.gz

Link refseq to VEP ensembl

ftp://ftp.ensembl.org/pub/release-75/variation/VEP/homo_sapiens_refseq_vep_75.tar.gz

ftp://ftp.ensembl.org/pub/release-75/variation/VEP/homo_sapiens_vep_75.tar.gz

3.1.1 SnpSift

a) Annotation with SnpSift

1 java -jar snpEff/SnpSift.jar dbnsfp -v snpEff/dbNSFP2.4.txt.gz P1_snp_chrA_combined.vcf > P1_snps_dbnsfp.vcf

b) Annotation dbSNP 141

2 java –jar snpEff/SnpSift.jar annotate dbSNP141.vcf P1_snps_dbnsfp.vcf > P1_snps_dbsnp.vcf

3.1.2 VEP ensembl (approximately 1h)

a) Annotation with VEP

3 perl variant_effect_predictor.pl --cache --everything --pick -i P1_snp_chrA_combined.vcf -o P1_annot_vep.vcf

Appendix B - Fisher Strand Bias calculation example

2 x 2 frequency table:

NLRC3	Strand (+)	Strand (-)	Sum
ref alleles	13	2	15
obs alleles	12	4	16
Sum	25	6	31

Contingency tables – all configurations with the same frequencies

Table A		
9	6	15
16	0	16
25	6	31

Table B		
10	5	15
15	1	16
25	6	31

Table C		
11	4	15
14	2	16
25	6	31

Table D		
12	3	15
13	3	16
25	6	31

Table E		
13	2	15
12	4	16
25	6	31

Table F		
14	1	15
11	5	16
25	6	31

Table G		
15	0	15
10	6	16
25	6	31

Fisher probability:

$$p = \frac{\binom{a+b}{a}\binom{c+d}{c}}{\binom{n}{a+c}} = \frac{(a+b)!\ (c+d)!\ (a+c)!\ (b+d)!}{a!\ \ b!\ \ c!\ \ d!\ \ n!}$$

Table A – Fisher probability = 0.0068

Table B – Fisher probability = 0.0653

Table C – Fisher probability = 0.222

Table D – Fisher probability = 0.346

Table E – Fisher probability = 0.260

Table F – Fisher probability = 0.0890

Table G – Fisher probability = 0.0109

p-value = 0.0068 + 0.0653 + 0.222 + 0.260 + 0.0890 + 0.0109 = 0.6539 (the probability for the Table D is not included because it is more probable than the observed frequency configuration)

Fisher strand bias = $-10\log_{10}0.6539 = 1.8446$

Appendix C –Results for the two models from our WES analysis

Table C.1. Variants obtained from the SNPs category from our bioinformatics pipeline (model 1). We applied three filters (as described in sub-chapter 3.6) and identified 16 variants. These are absent from the unaffected (ID IV.1-080002) but present in all five affected individuals. Also, these variants are not present in public databases.

Chr	Position (bp)	Ref allele	Obs allele	Gene	CDS change	Variant classification	SIFT score	PolyPhen-2 score
2	1.457.407	T	C	*TPO*	c.483-59T>C	Intron	-	-
2	171.702.584	C	T	*GAD1*	c.1002+11C>T	Intron	-	-
6	32.609.813	T	C	*HLA-DQA1*	c.396T>C	Synonymous	-	-
6	32.709.346	T	C	*HLA-DQA2*	c.82+44T>C	Intron	-	-
6	32.709.370	T	C	*HLA-DQA2*	c.82+68T>C	Intron	-	-
8	145.171.113	C	T	*KIAA1875*	c.4786C>T	NSV	0.000	0.996
9	85.622.564	T	C	*RASEF*	c.960-144A>G	Intron	-	-
12	121.097.538	A	G	*CABP1*	c.655-143A>G	Intron	-	-
16	2.105.335	C	G	*TSC2*	c.482-68C>G	Intron	-	-
16	3.627.162	G	A	*NLRC3*	c.53C>T	NSV	0.920	0.000
16	4.245.598	A	G	*SRL*	c.566T>C	NSV	0.000	0.948
16	4.513.886	G	A	*NMRAL1*	c.530-35C>T	Intron	-	-
16	1.330.7046	G	A	*SHISA9*	c.848-64G>A	Intron	-	-
17	43.924.412	T	C	*SPPL2C*	c.*85T>C	3'UTR	-	-
21	9.912.194	T	C	*TEKT4P2*	n.130-2917A>G	Intron	-	-
21	9.914.197	C	T	*TEKT4P2*	n.130-4920G>A	Intron	-	-

Chr: Chromosome; Ref: Reference; Obs: Observed; CDS change: Coding DNA sequence change; NSV: Non-synonymous Variants.

Table C.2. Variants obtained from the InDels category from our bioinformatics pipeline (model 1). We applied three filters (as described in sub-chapter 3.6) and identified one variant. These are absent from the unaffected (ID IV.1-080002) but present in all five affected individuals. Also, these variants are not present in public databases.

Chr	Position (bp)	Ref allele	Obs allele	Gene	CDS change	Variant classification	SIFT score	PolyPhen-2 score
16	2.115.413	CAAAAAAA AAAAAAAA	C	*TSC2*	c.1600-106_1600-92delAAAAAAAAAAAAAAA	Intron	-	-

Chr: Chromosome; Ref: Reference; Obs: Observed; CDS change: Coding DNA sequence change.

Table C.3. Variants obtained from the SNPs category from our bioinformatics pipeline (model 2). We applied three filters (as described in sub-chapter 3.6) and identified eight variants. These are absent from the two unaffected relatives (ID IV.1-080002 and IV.3-090095) but present in all four affected individuals. Also, these variants are not present in public databases.

Chr	Position (bp)	Ref allele	Obs allele	Gene	CDS change	Variant classification	SIFT score	PolyPhen-2 score
1	26.863.294	G	A	*RPS6KA1*	c.64-122G>A	Intron	-	-
1	31.896.668	G	A	*SERINC2*	c.195G>A	NSV	0.030	0.068
4	69.540.024	T	C	*UGT2B15*	-	Upstream	-	-
12	58.218.301	A	G	*CTDSP2*	c.412-199T>C	Intron	-	-
19	32.130.944	T	C	-	-	Intergenic	-	-
19	35.530.272	C	G	*HPN*	-	Upstream	-	-
19	36.018.347	G	T	*SBSN*	c.837C>A	Synonymous	-	-
21	9.913.829	C	T	*TEKT4P2*	n.130-4552G>A	Intron	-	-

Chr: Chromosome; Ref: Reference; Obs: Observed; CDS change: Coding DNA sequence change; NSV: Non-synonymous Variants.

Table C.4. Variants obtained from the InDels category from our bioinformatics pipeline (model 2). We applied three filters (as described in sub-chapter 3.6) and identified two variants. These are absent from the two unaffected relatives (ID IV.1-080002 and IV.3-090095) but present in all four affected individuals. Also, these variants are not present in public databases.

Chr	Position (bp)	Ref allele	Obs allele	Gene	CDS change	Variant classification	SIFT score	PolyPhen-2 score
1	6.529.182	TTCCTCC	T	*PLEKHG5*	c.2400_2405delGGAGGA	NFD	-	-
12	6.909.380	TG	T	*CD4*	c.49+28delG	Intron	-	-

Chr: Chromosome; Ref: Reference; Obs: Observed; CDS change: Coding DNA sequence change; NFD: Non-Frameshift Deletion.

Appendix D - PCR conditions

Variants	*Stepdown* PCR conditions
c.C8800T (G/A) and c.C871T (G/A)	25 cycles 94°/30 seconds 70°/30 seconds (-1° C/cycle) 72°/60 seconds 30 cycles 94°/10 seconds 55°/45 seconds 72°/60 seconds 72°/7 minutes 4°/∞
c.T1933C (T/C) and c.2163_2168del (TTCCTCC/T)	20 cycles 94°/30 seconds 70°/30 seconds (-1° C/cycle) 72°/60 seconds 30 cycles 94°/10 seconds 55°/45 seconds 72°/60 seconds 72°/7 minutes 4°/∞
c.C2264T (G/A)	20 cycles 94°/30 seconds 70°/30 seconds (-0.5° C/cycle) 72°/60 seconds 30 cycles 94°/10 seconds 60°/45 seconds 72°/60 seconds 72°/7 minutes 4°/∞
c.C4786T (C/T)	20 cycles 94°/30 seconds 65°/30 seconds (-0.5° C/cycle) 72°/60 seconds 30 cycles 94°/10 seconds 53°/45 seconds 72°/60 seconds 72°/7 minutes 4°/∞
c.C53T (G/A)	20 cycles 94°/30 seconds 70°/30 seconds (-1° C/cycle) 72°/30 seconds 30 cycles 94°/10 seconds 55°/45 seconds 72°/60 seconds 72°/7 minutes 4°/∞
c.T566C (A/G)	20 cycles 94°/30 seconds 65°/30 seconds (-0.5° C/cycle) 72°/60 seconds 30 cycles 94°/10 seconds 52°/45 seconds 72°/60 seconds 72°/7 minutes 4°/∞
c.G180A (G/A)	20 cycles 94°/30 seconds 70°/30 seconds (-0.5° C/cycle) 72°/60 seconds 42 cycles 94°/10 seconds 50°/45 seconds 72°/60 seconds 72°/7 minutes 4°/∞

Appendix E - Ethics committee approval

INSTITUTO PORTUGUÊS DE REUMATOLOGIA

Exmo. Senhor
Dr. José Vaz Patto
Investigador principal
I.P.R

Lisboa, 2 de Maio de 2008

Exmo. Senhor,

A Comissão de Ética para a Saúde do Instituto Português de Reumatologia, vem informar V. Exa. que aprovou na sua reunião do dia 2 de Maio de 2008, a realização do estudo "Estudo de factores genéticos numa família portuguesa com Doença Óssea de Paget", no Instituto de Reumatologia, mas aguarda-se o envio dos curriculuns respectivos.

Com os meus cumprimentos.

A Comissão de Ética

PP/MP

Appendix F – Quality control

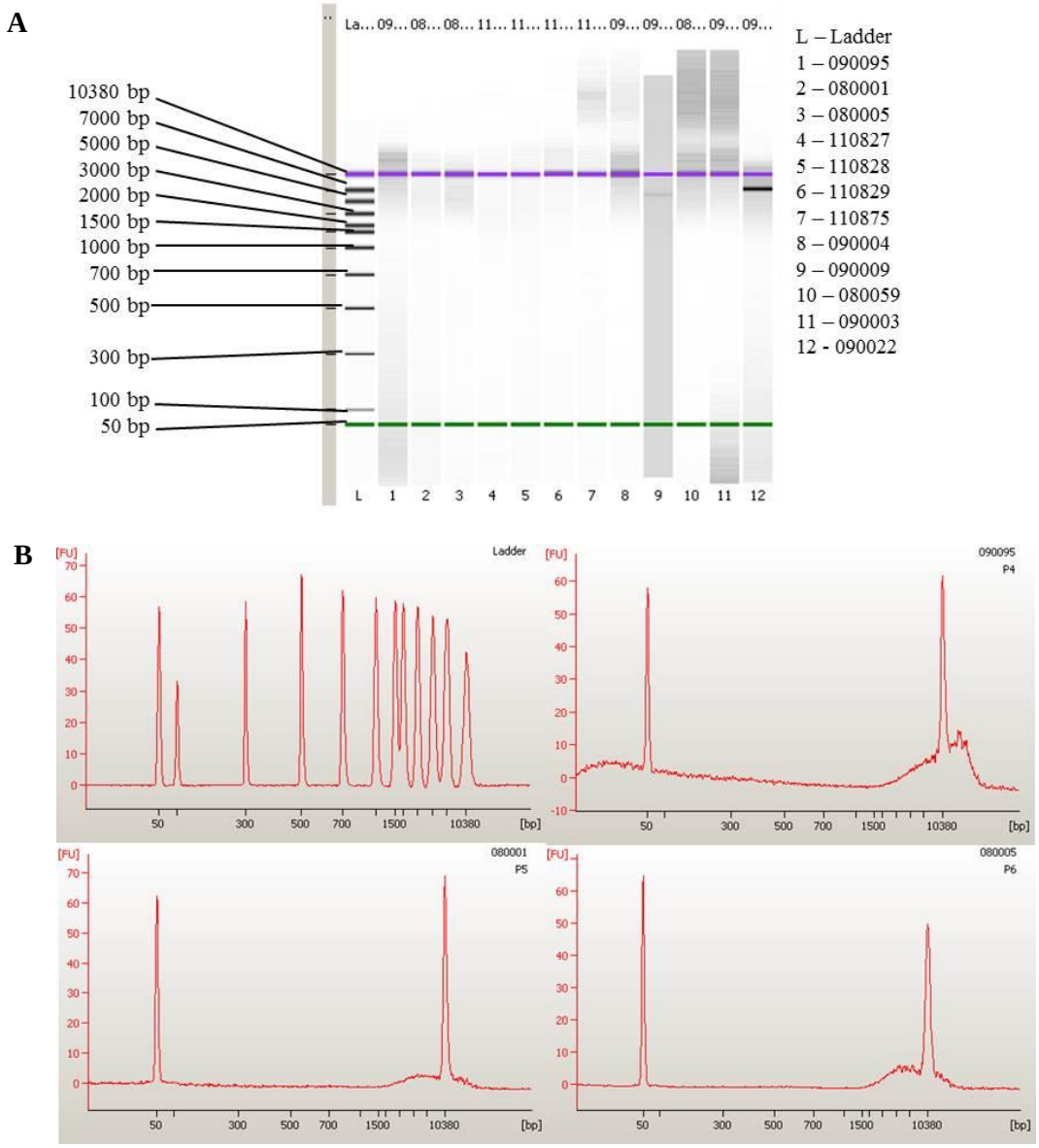
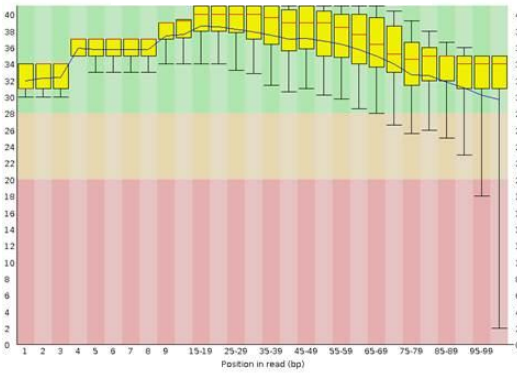


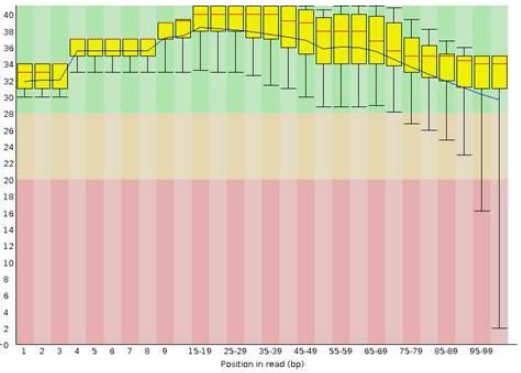
Figure F.1. Bioanalyzer results. A – After running the three PDB samples and nine controls, in the DNA LabChip, a gel-like image of the ladder and all DNA samples was obtained. **B –** Electropherograms of the ladder and three PDB samples. These graphs depicted the fluorescence units (FU) per base pair. Two markers bracketing the overall sizing range were run with each sample. The two peaks represented in the electropherograms are the "lower" (50 bp) and "upper" (10380 bp) markers (internal standards).

Appendix G – Sequence Quality graphs

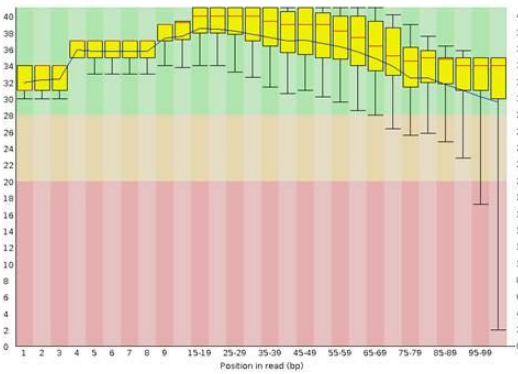
A1



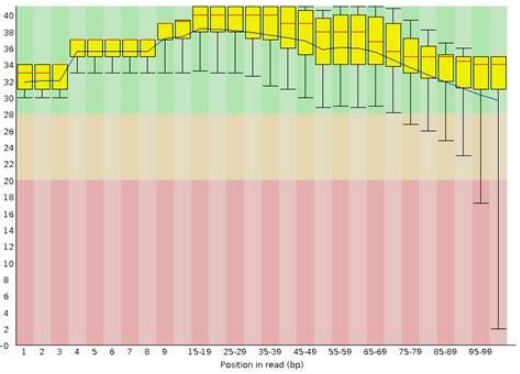
A2



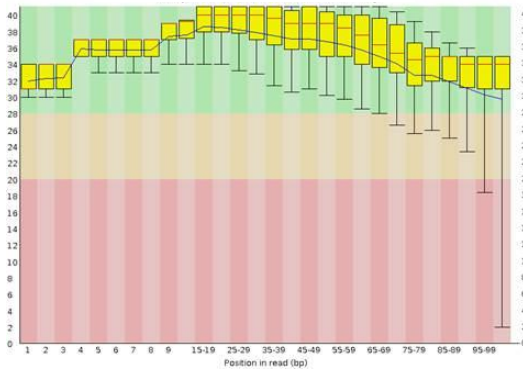
B1



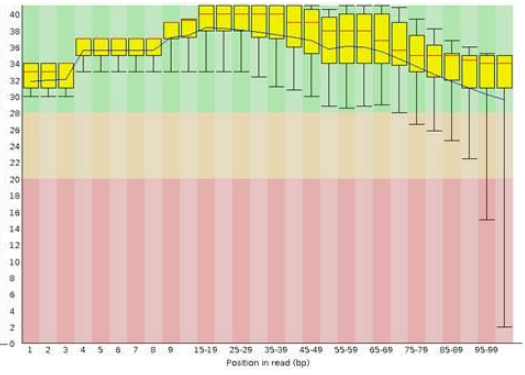
B2



C1



C2



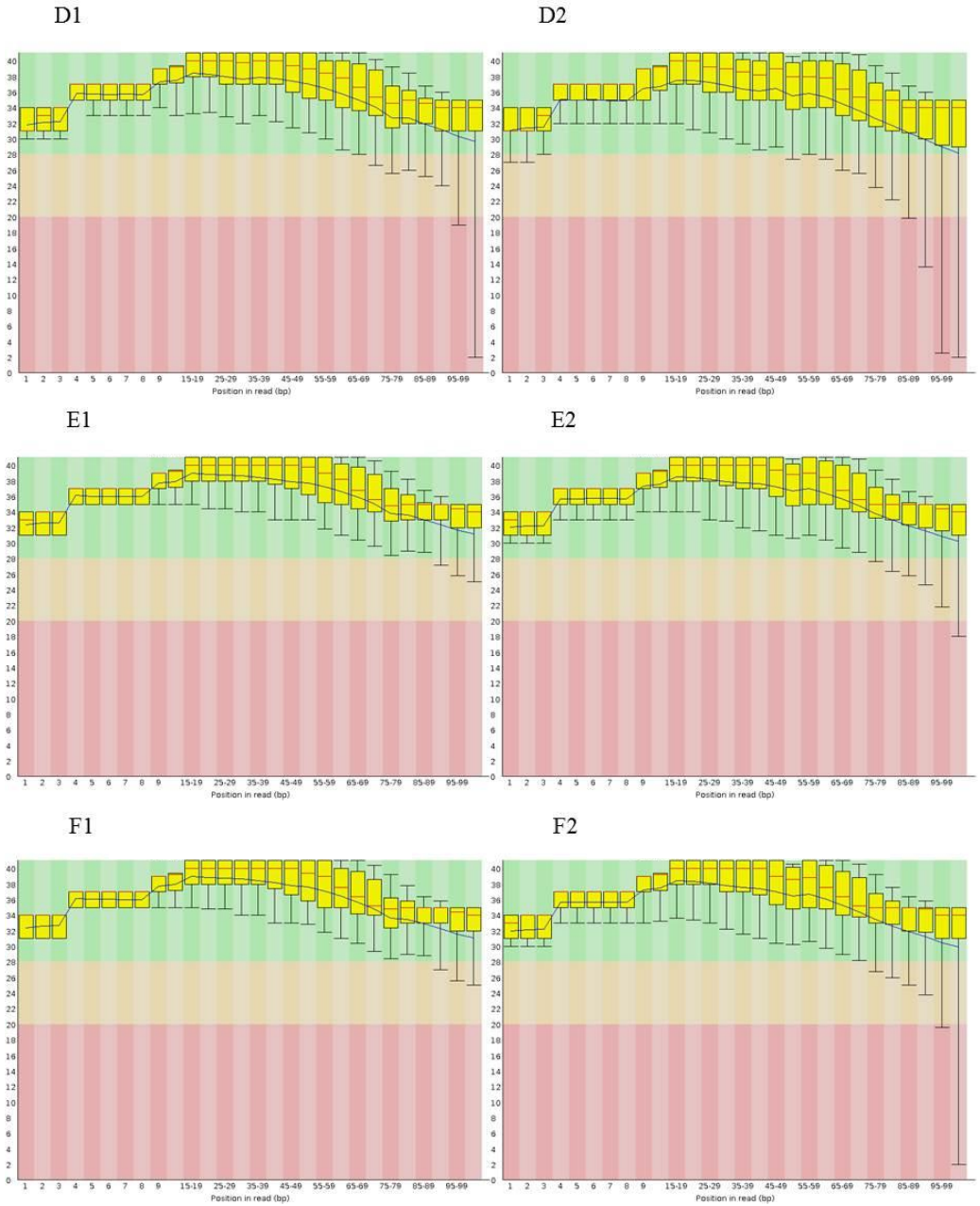
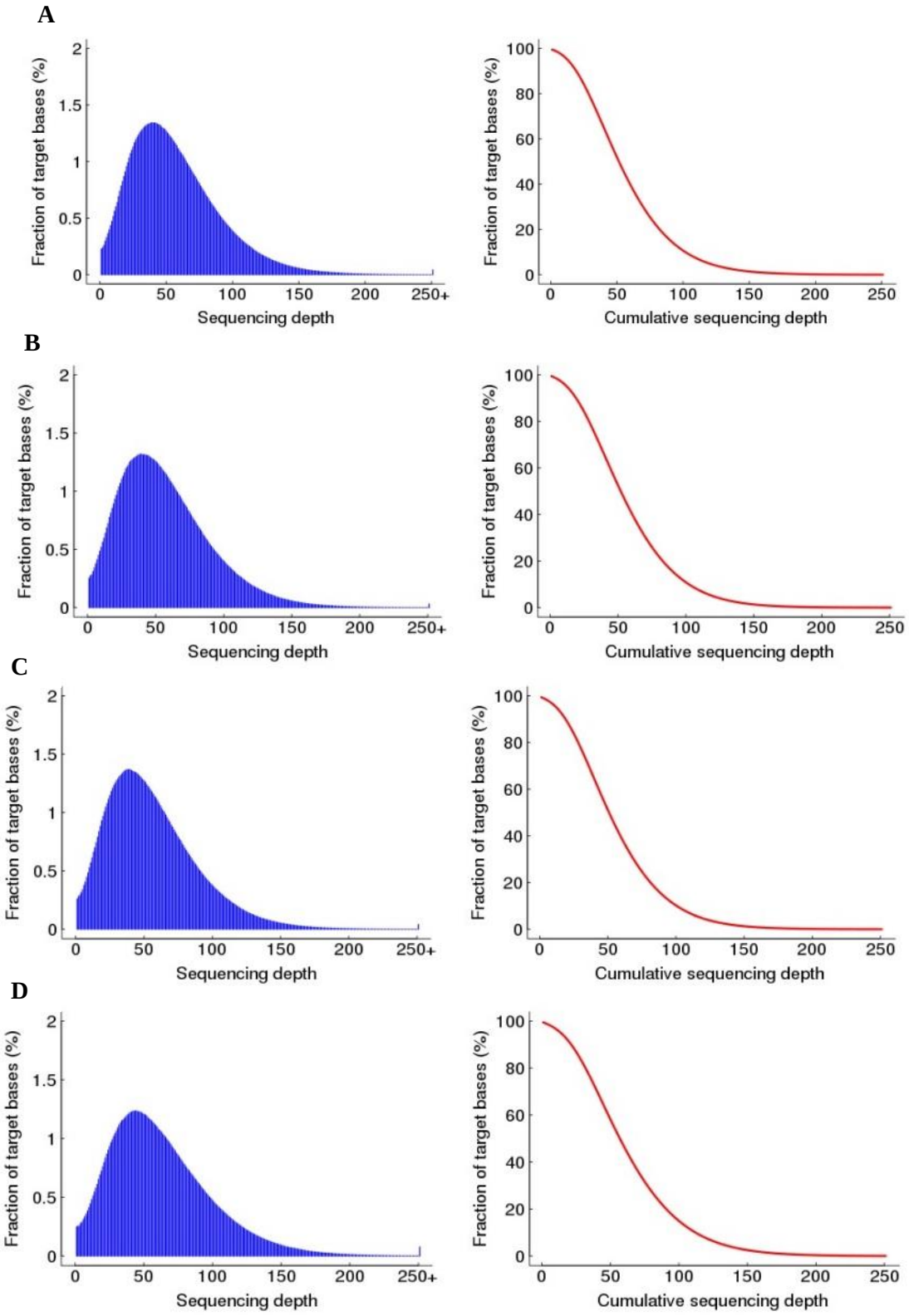


Figure G.1. Sequence quality graphs per base for the six individuals exome-sequenced using the FASTQ1 (A1 through F1) and the FASTQ2 (A2 through F2) files. A1 and A2 – III.2-080004; **B1 and B2 –** IV.1-080002; **C1 and C2 –** IV.9-090044; **D1 and D2 –** IV.3-090095; **E1 and E2 –** III.4-080001; **F1 and F2 –** III.6-080005. The x-axis shows the position in the read (in base pairs) and the y-axis shows the quality score across all bases. The graph is divided into three colours: green, orange and red represent the calls with good, reasonable and bad quality, respectively. The higher the quality score the better the base call. The quality of the calls typically degrades as the run progresses. At each position a whisker box-plot was drawn. The yellow box represents the inter-quartile range of each position (from 25 to 75%). The upper and lower whiskers represent the 10% and 90% points. Central red line is the median value and the blue line represents the mean quality.


A
Fraction of target bases (%)
Sequencing depth
Cumulative sequencing depth
B
C
D

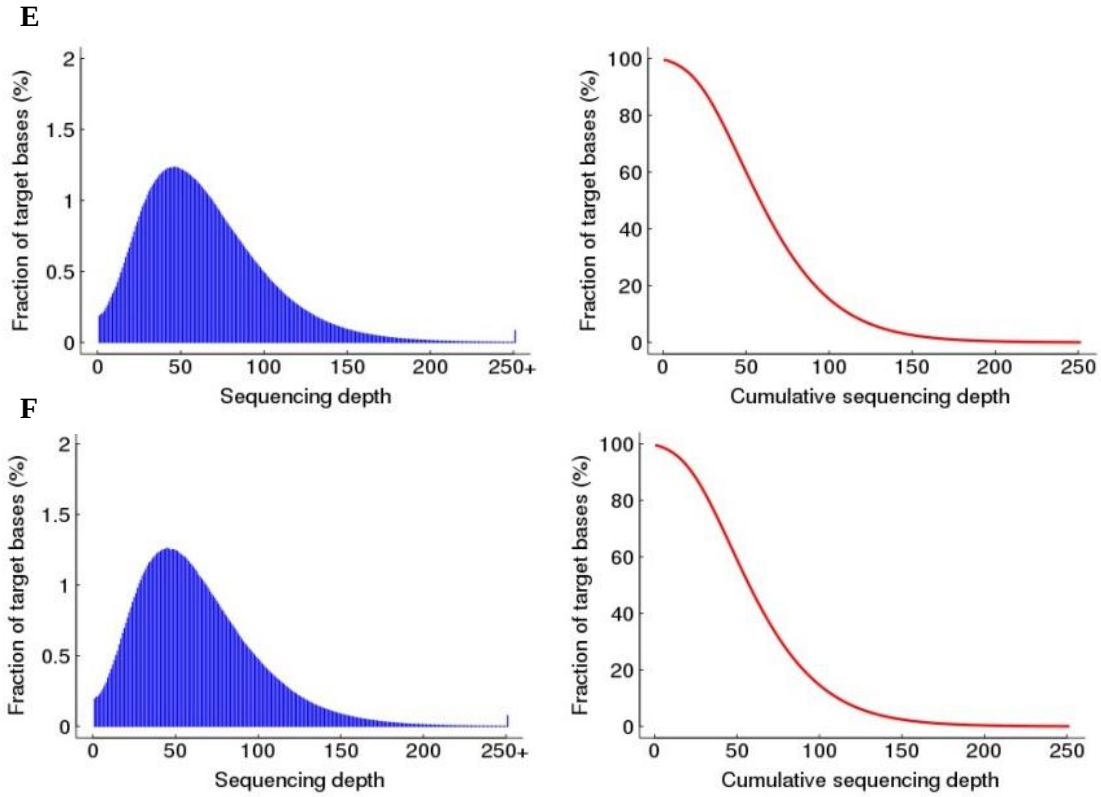
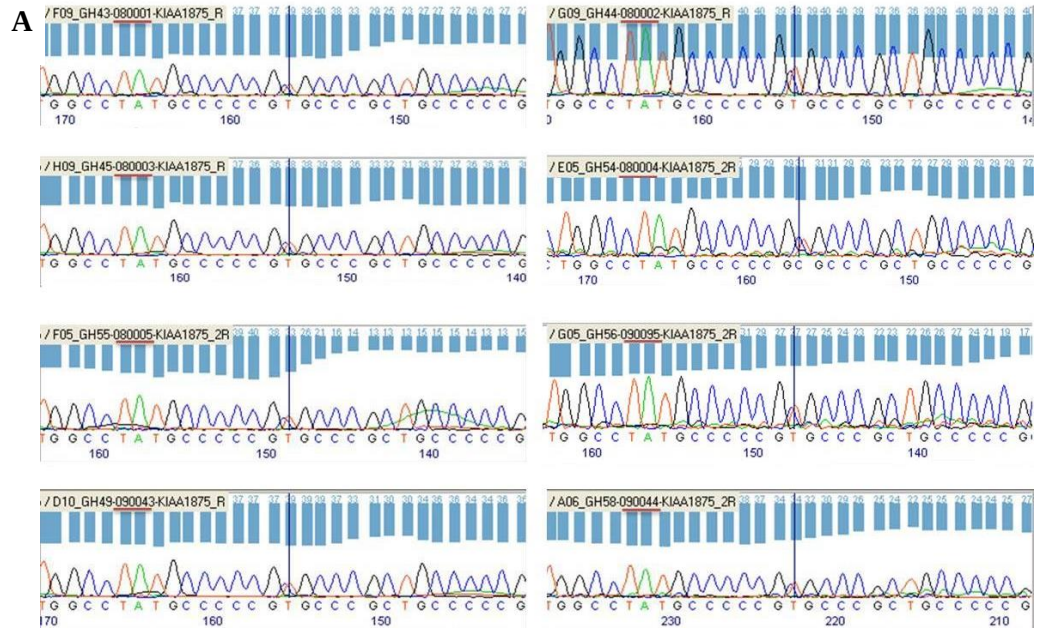


Figure G.2. Sequencing and cumulative sequencing depth graphs for the six individuals exome-sequenced. A: III.2-080004; **B:** IV.1-080002; **C:** IV.9-090044; **D:** IV.3-090095; **E:** III.4-080001; **F:** III.6-080005).

Appendix H – Sanger Sequencing chromatograms

T G G C C T A T G C C C C C G [C/T] G C C C C G C T G C C C C C G – Family 1



T G G C C T A T G C C C C C G [C/T] G C C C C G C T G C C C C C G – Family 2

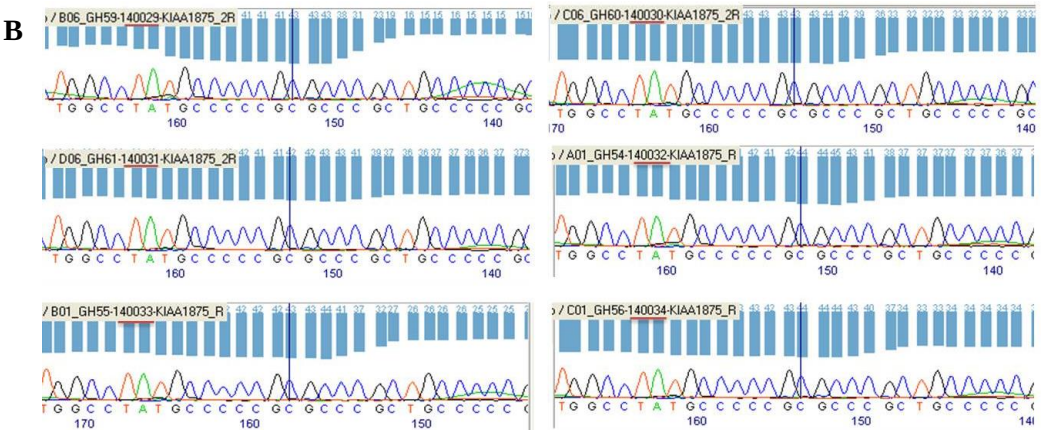
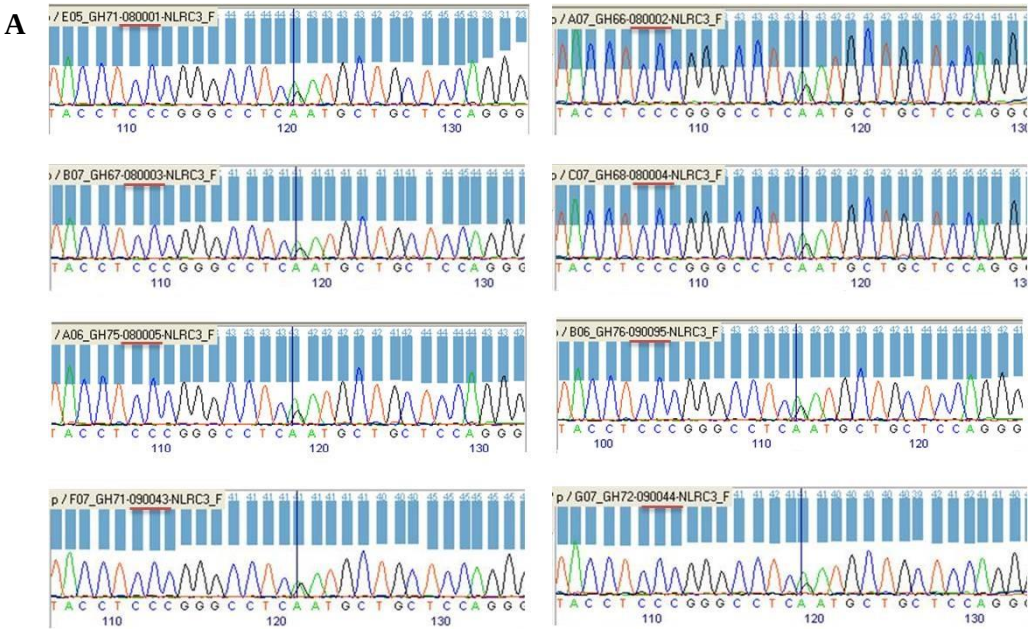


Figure H.1. Chromatograms for the c.C4786T variant (reverse strand). A – Family 1: the mutated T allele for the c.C4786T variant was detected in patients (III.1-080003, III.2-080004, III.4-080001, III.6-080005, IV.3-090095 and IV.9-090044), and in the controls (III.8-090043 and IV.1-080002). **B –** Family 2: the ancestral C allele for the c.C4786T variant was detected in patients (II.2-140030 and II.3-140029), in the unclear individuals (III.1-140031 and III.2-140034) and in controls (II.1-140033 and IV.1-140032).

A C C T C C C G G G C C T C [G /A] A T G C T G G C T C C A G G – Family 1



A C C T C C C G G G C C T C [G /A] A T G C T G G C T C C A G G – Family 2

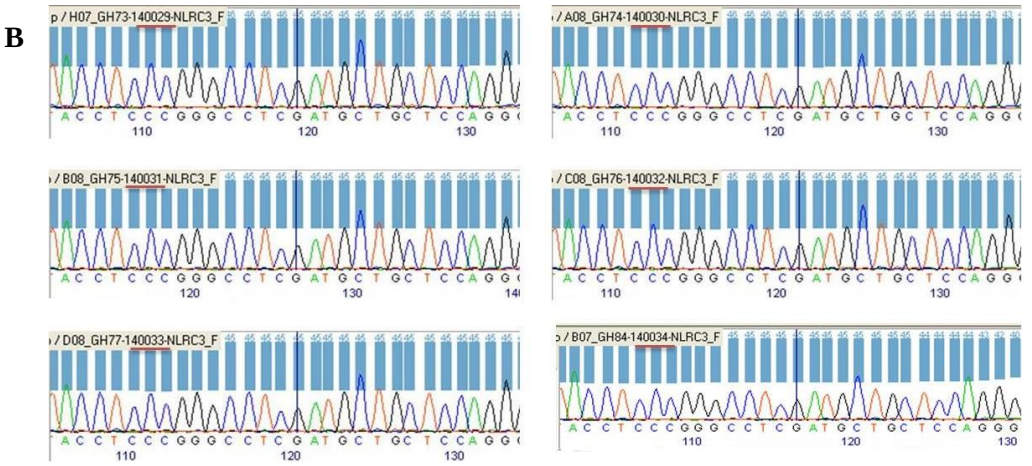
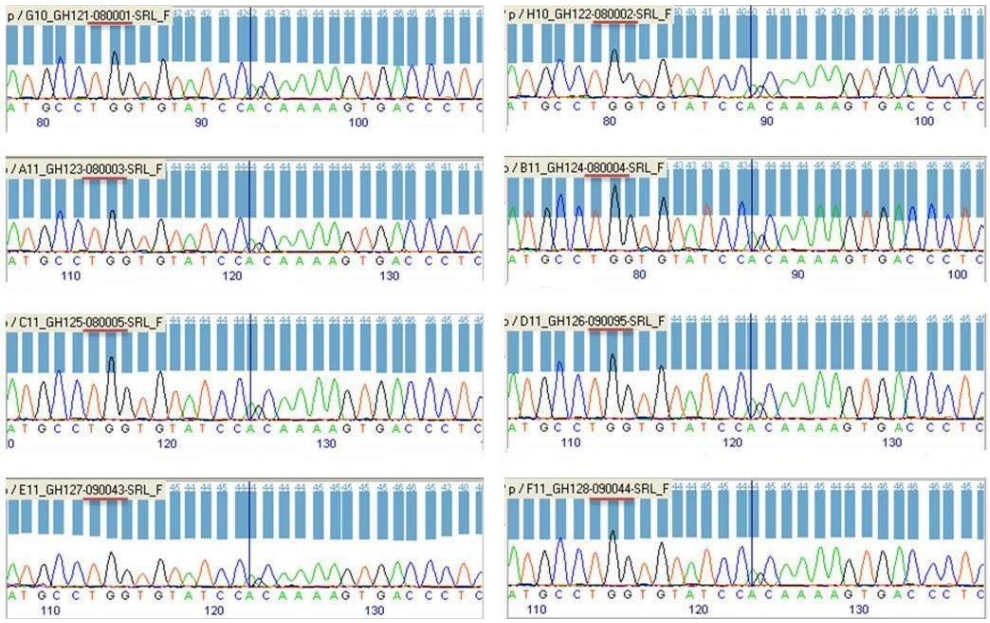


Figure H.2. Chromatograms for the c.C53T variant (forward strand). A – Family 1: the mutated A allele for the c.C53T variant was detected in patients (III.1-080003, III.2-080004, III.4-080001, III.6-080005, IV.3-090095 and IV.9-090044), and in the controls (III.8-090043 and IV.1-080002). **B –** Family 2: the ancestral G allele for the c.C53T variant was detected in patients (II.2-140030 and II.3-140029), in the unclear individuals (III.1-140031 and III.2-140034) and in controls (II.1-140033 and IV.1-140032).

A T G C C T G G T G T A T C C [A/G] C A A A A A G T G A C C C T C – Family 1

A



A T G C C T G G T G T A T C C [A/G] C A A A A A G T G A C C C T C – Family 2

B

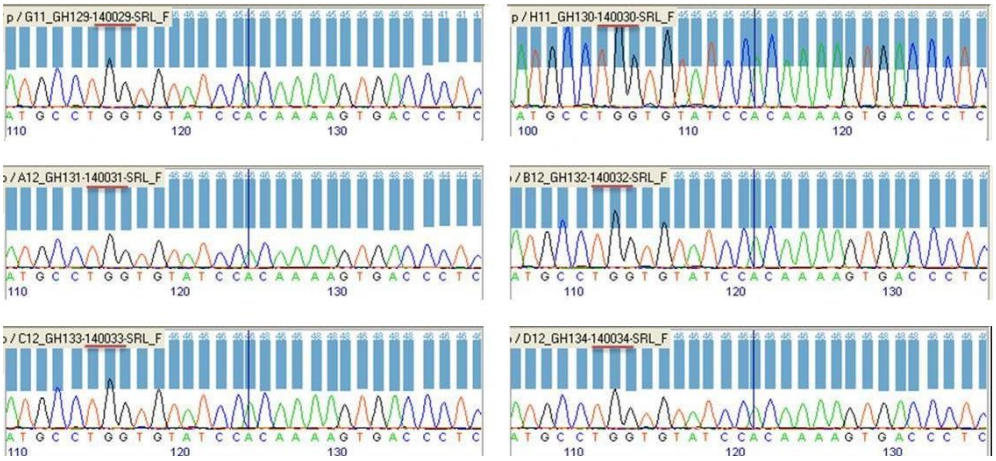
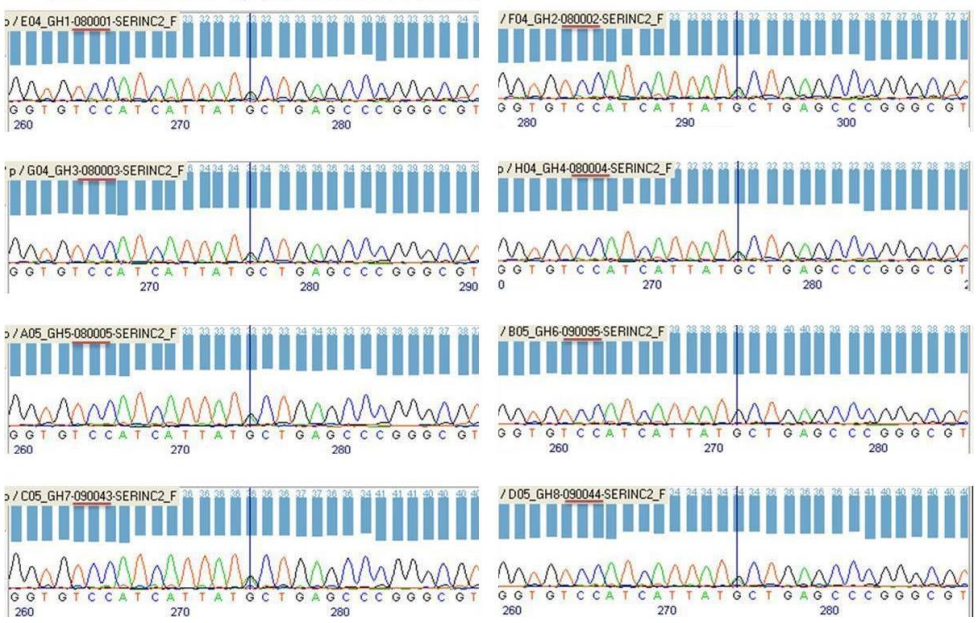


Figure H.3. Chromatograms for the c.T566C variant (forward strand). A – Family 1: the mutated G allele for the c.T566C variant was detected in patients (III.1-080003, III.2-080004, III.4-080001, III.6-080005, IV.3-090095 and IV.9-090044), and in the controls (III.8-090043 and IV.1-080002). **B –** Family 2: the ancestral A allele for the c.T566C variant was detected in patients (II.2-140030 and II.3-140029), in unclear individuals (III.1-140031 and III.2-140034) and in controls (II.1-140033 and IV.1-140032).

G G T G T C C A T C A T T A T [G/A] C T G A G C C C G G G C G T – Family 1

A



G G T G T C C A T C A T T A T [G/A] C T G A G C C C G G G C G T – Family 2

B

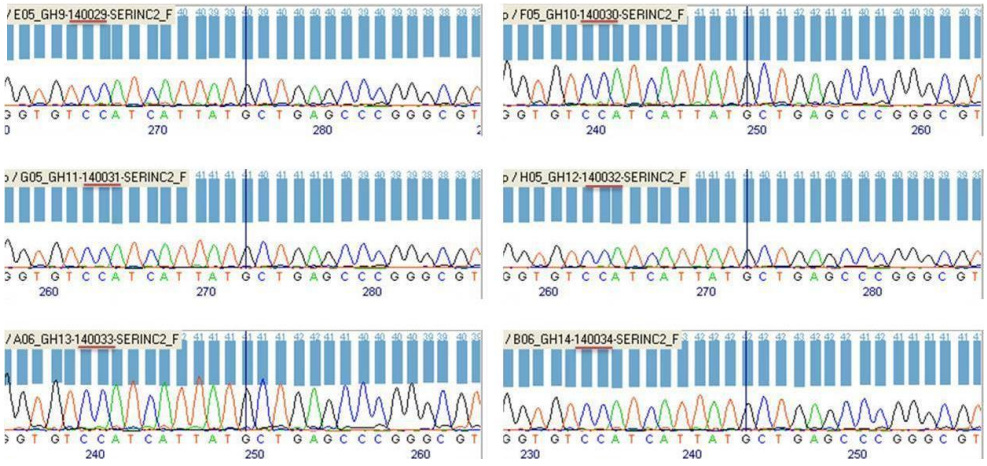
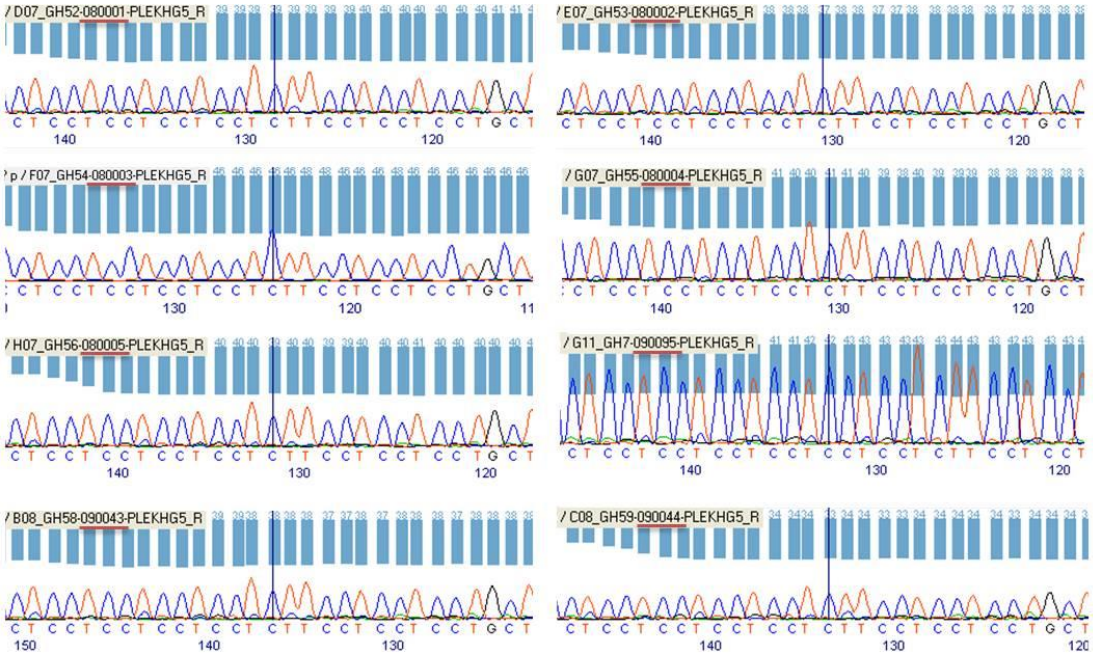


Figure H.4. Chromatograms for the c.G180A variant (forward strand). A – Family1: the mutated A allele for the c.G180A variant was detected in patients (III.1-080003, III.2-080004, III.4-080001, III.6-080005 and IV.9-090044), and in the controls (III.8-090043 and IV.1-080002). The variant is absent in the unclear individual IV.3-090095. **B –** Family 2: the ancestral G allele for the c.G180A variant was detected in patients (II.2-140030 and II.3-140029), in the unclear individuals (III.1-140031 and III.2-140034) and in controls (II.1-140033 and IV.1-140032).

C T C C T C C T C C T C C T [C C T C C T C / C] T T C C T C C T C C T G C T – Family 1



C C T T C C T C C T C C T C C T C C T C C T C C T [C C T C C T C / C] T T C – Family 2

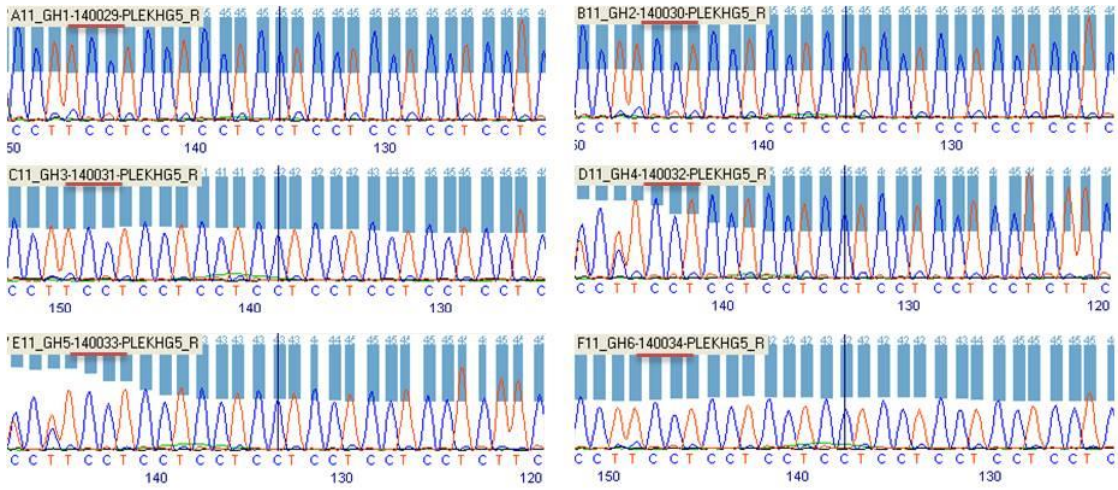


Figure H.5. Chromatograms for the c.2163_2168del variant (reverse strand). A – Family 1: the deletion CTCCTC for the 6.529.184/6.529.190 bp position in chromosome 1 was detected in patients (III.1-080003, III.2-080004, III.4-080001, III.6-080005 and IV.9-090044), and in the controls (III.8-090043 and IV.1-080002). The deletion is absent in the unclear individual IV.3-090095. **B –** Family 2: the deletion CTC for the 6.529.187/6.529.190 bp position in chromosome 1 was detected in the controls (II.1-140033 and IV.1-140032), and was absent in patients (II.2-140030 and II.3-140029) and in the unclear individuals (III.1-140031 and III.2-140034).

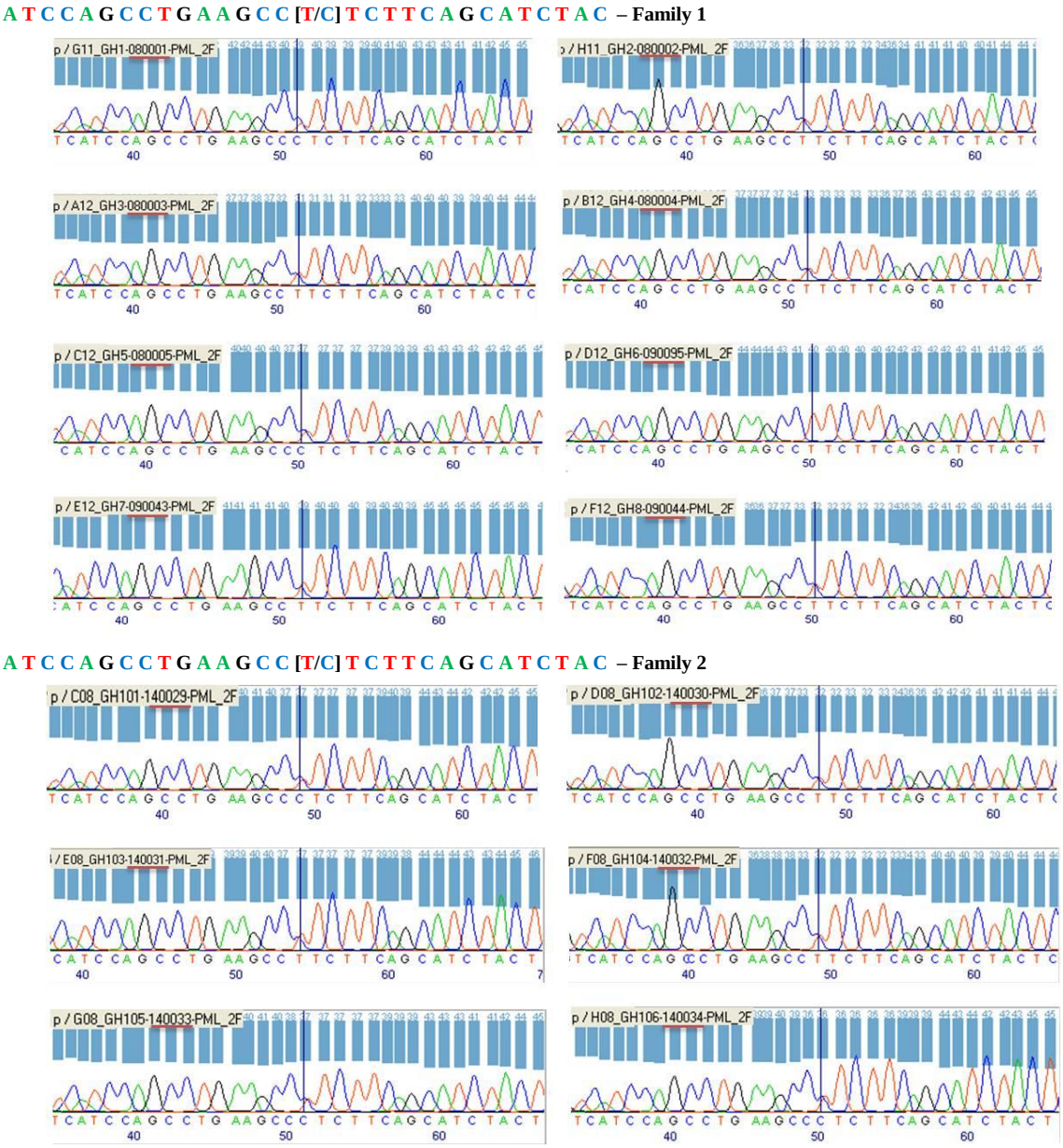


Figure H.6. Chromatograms for the c.T1933C variant (forward strand). A –the mutated C allele for c.T1933C variant was detected in patients (III.1-080003, III.2-080004, III.4-080001, III.6-080005 and IV.9-090044), and in controls (III.8-090043 and IV.1-080002) from family 1. The variant is absent in the unclear individual IV.3-090095 from family 1. **B** –the mutated C allele, for c.T1933C variant was detected in patients (II.2-140030 and II.3-140029), in the unclear individuals (III.1-140031 and III.2-140034) and in both controls (II.1-140033 and IV.1-140032) from family 2.

Appendix I – Variants in genes associated with PDB

Table I.1. Exonic, regulatory, intronic, intergenic and splicing variants in PDB-associated genes. Variants present at least one individual from family 1.

Individual	Variant classification	Gene	CDS change	dbSNP	1000 GP (MAF)	SIFT score	PolyPhen-2 score	Chr	Start position (bp)	Ref allele	Obs allele
P2, C, U	Upstream		-	rs1999714	0.556	-	-	1	110.450.033	G	T
U	Upstream		-	rs1999713	0.551	-	-	1	110.450.177	T	C
U	Intron		c.162+65G>A	.	-	-	-	1	110.457.068	G	A
P1, P3, P4, C, U	Intron		c.226-142C>T	rs333967	0.699	-	-	1	110.459.773	C	T
C	Intron		c.396+1569G>A	rs333969	0.001	-	-	1	110.461.654	G	A
P2, U	Synonymous	*CSF1*	c.1095C>A	rs333970	0.553	-	-	1	110.466.338	C	A
P1, P2, P3, P4, C, U	NSV		c.1223T>C	rs1058885	0.574	0.300	0.011	1	110.466.466	T	C
P1, P2, P3, P4, C, U	NSV		c.1466T>C	rs333971	0.002	0.970	0.003	1	110.466.709	T	C
P1, P2, P3, P4, C, U	Intron		c.1623-24A>G	rs333972	0.002	-	-	1	110.467.745	A	G
P2	Splice region		c.*13+7T>A	rs41313284	0.993	-	-	1	110.467.831	T	A
P1, P2, P3, P4, C, U	Intron		c.*13+185C>A	rs3138072	0.905	-	-	1	110.468.009	C	A
P1	Upstream		-	rs6776158	0.411	-	-	3	121.901.849	G	A
U	Intron		c.-243+6512A>G	rs73858157	0.855	-	-	3	121.909.888	A	G
U	Intron		c.-243+7319C>G	rs1553308	0.254	-	-	3	121.910.695	C	G
U	Intron		c.-243+7330A>G	rs1553309	0.662	-	-	3	121.910.706	A	G
C	Intron		c.-243+10478_-243+10479insCTC	rs111507015	-	-	-	3	121.913.854	G	GCTC
U	Intron	*CaSR*	c.-243+19950G>A	rs62271400	0.628	-	-	3	121.923.326	G	A
U	Intron		c.-243+32891A>G	rs11721042	0.321	-	-	3	121.936.267	A	G
U	Intron		c.-243+32902T>C	rs11717321	0.322	-	-	3	121.936.278	T	C
U	Intron		c.-243+34086_-243+34088delAGA	rs34710605	0.335	-	-	3	121.937.461	TAGA	T
C	Intron		c.-242-3948T>G	rs2173961	0.369	-	-	3	121.968.847	T	G
C, U	Intron		c.-242-1436_-242-1433delAGAA	rs141055482	0.375	-	-	3	121.971.358	GAGAA	G

C	Intron		c.-242-1283G>T	rs937625	0.175	-	-	3	121.971.512	G	T
P1, P2, P3, P4, C, U	Intron		c.492+19G>A	rs9869985	0.078	-	-	3	121.976.253	G	A
P1, P2, U	Intron		c.493-133T>C	rs3749207	0.465	-	-	3	121.980.242	T	C
P2	Intron		c.493-91C>T	rs3749208	0.666	-	-	3	121.980.284	C	T
U	Intron		c.1377+2546C>T	rs36060529	0.799	-	-	3	121.983.805	C	T
P2	Intron		c.1378-1185C>T	rs35780274	0.747	-	-	3	121.993.474	C	T
P1, P2, P3, U	Intron	*CaSR*	c.1609-59C>T	rs4678174	0.468	-	-	3	122.000.871	C	T
P3, P4	Intron		c.1762+16T>C	rs2270916	0.821	-	-	3	122.001.099	T	C
P1, U	Intron		c.1762+163C>T	rs2270917	0.468	-	-	3	122.001.246	C	T
P1, P2, P3, P4, C, U	Synonymous		c.2274G>C	rs2036400	0.025	-	-	3	122.003.045	G	C
P2	NSV		c.2986G>T	rs386545639	0.924	0.170	0.012	3	122.003.757	G	T
P1, P2, P3, P4, C, U	NSV		c.3061G>C	rs1801726	0.077	0.910	0.000	3	122.003.832	G	C
P1, P2, P3, P4, U	3'UTR		c.*60A>T	rs4677948	0.078	-	-	3	122.004.098	A	T
U	Intron		c.-48+2228A>C	rs10516140	0.259	-	-	5	179.240.910	A	C
U	Upstream		-	rs172057	0.008	-	-	5	179.245.959	C	T
U	Intron		c.673+152T>G	rs55993594	0.440	-	-	5	179.251.475	T	G
P3	Intron		c.755-23G>A	rs386562270	0.599	-	-	5	179.260.009	G	A
P3, U	Synonymous	*SQSTM1*	c.876C>T	rs4935	0.317	-	-	5	179.260.153	C	T
P3, U	Synonymous		c.936G>A	rs4797	0.420	-	-	5	179.260.213	G	A
P3	Intron		c.970-109C>G	rs2241350	0.859	-	-	5	179.260.478	C	G
P1, P2, P3, P4	Intron		c.970-93G>A	rs155787	0.388	-	-	5	179.260.494	G	A
U	3'UTR		c.*1322G>T	rs1065154	0.306	-	-	5	179.264.915	G	T
U	Intron		n.73+1350A>G	rs851969	0.857	-	-	6	151.979.248	A	G
U	Intron		n.73+8791T>C	rs1293959	0.000	-	-	6	151.986.689	T	C
U	Intron	*ESR1*	n.73+8837G>T	rs1293958	0.000	-	-	6	151.986.735	G	T
P1	Intron		n.73+9908C>A	rs7745737	0.748	-	-	6	151.987.806	C	A
U	Intron		c.-71+41215G>A	rs1999806	0.609	-	-	6	152.064.355	G	A
U	Intron		c.-71+46651A>G	rs2982575	0.607	-	-	6	152.069.791	A	G

U	Intron		c.-70-31281C>G	rs2504068	0.376	-	-	6	152.097.697	C	G
P2, P4, C	Synonymous		c.30T>C	rs386556657	0.568	-	-	6	152.129.077	T	C
U	Intron		c.453-8005A>T	rs827424	0.179	-	-	6	152.155.727	A	T
P1, P2, P3, P4, C, U	Synonymous		c.729T>C	rs4986934	0.018	-	-	6	152.201.875	T	C
P1, P2, P3, P4	Intron	*ESR1*	c.760+101T>C	rs3757323	0.558	-	-	6	152.202.007	T	C
P2	Intron		c.760+22195A>G	rs12215922	0.841	-	-	6	152.224.101	A	G
P2	Intron		c.760+27539T>C	rs1514347	0.273	-	-	6	152.229.445	T	C
P1, P2	Intron		c.761-26_761-25insT	rs55740371	0.512	-	-	6	152.265.282	A	AT
P1, P2, P3, P4, C, U	Synonymous		c.975G>C	rs1801132	0.262	-	-	6	152.265.522	G	C
C, U	Synonymous		c.1782G>A	rs2228480	0.821	-	-	6	152.420.095	G	A
P3	Upstream		-	rs4728356	0.015	-	-	7	135.240.927	C	T
C	Intron		c.28+27T>C	rs117892953	0.988	-	-	7	135.242.747	T	C
P1, P3, P4, C	Intron		c.28+190C>T	rs6961420	0.494	-	-	7	135.242.910	C	T
P1, P3, P4, C	Intron		c.649-11T>C	rs10252250	0.659	-	-	7	135.262.533	T	C
P1	Intron		c.1218+819A>G	rs7781687	0.001	-	-	7	135.270.574	A	G
P1, P3, P4, C	Intron		c.1219-53T>C	rs10271506	0.758	-	-	7	135.272.270	T	C
P1, P2, P3, P4, C, U	Intron		c.1624+137G>A	rs4410839	0.000	-	-	7	135.276.485	G	A
P1, P3, P4, C	Synonymous		c.1851A>C	rs7800214	0.653	-	-	7	135.279.315	A	C
P1, P2, P4, C, U	Intron	*NUP205*	c.2374+71T>C	rs4296989	0.009	-	-	7	135.285.788	T	C
U	Intron		c.3310+307G>A	rs4316099	0.016	-	-	7	135.299.328	G	A
P1, P2, P3, P4, C, U	NSV		c.4066G>C	rs7810767	0.014	0.730	0.001	7	135.304.273	G	C
P1, P3, P4, C, U	Intron		c.4671+102A>G	rs6972359	0.663	-	-	7	135.310.205	A	G
P1, P2, P3, P4, C, U	Intron		c.4672-128G>A	rs10085457	0.728	-	-	7	135.310.860	G	A
P2	Intron		c.4793+11G>A	rs7810260	0.586	-	-	7	135.311.120	G	A
P1, P2, P3, P4, C, U	Intron		c.4793+58G>A	rs10234309	0.282	-	-	7	135.311.167	G	A
P1, P2, P3, P4, C, U	Splice region		c.5559+4C>T	rs10260691	0.720	-	-	7	135.328.110	C	T
P2, P3, P4, U	Intron		c.5813-77C>A	rs62479523	0.885	-	-	7	135.330.829	C	A

P4, U	Splice region		c.817+8A>C	rs7844539	0.935	-	-	8	119.938.725	T	G
P4, U	Synonymous		c.768A>G	rs2228568	0.928	-	-	8	119.938.782	T	C
P4, U	Intron		c.592+55_592+56delCT	rs10554146	0.822	-	-	8	119.940.920	CAG	C
P1, P2, P3, P4, C, U	Splice region		c.401-5T>C	rs3134046	0.084	-	-	8	119.941.173	A	G
P3	Intron		c.401-109T>C	rs4876869	0.677	-	-	8	119.941.277	A	G
U	Intron	*TNFRSF11B*	c.401-1477C>T	rs11573916	0.929	-	-	8	119.942.645	G	A
P4, U	Splice region		c.400+4C>T	rs1564858	0.928	-	-	8	119.945.166	G	A
U	Intron		c.30+4374G>A	rs3134063	0.408	-	-	8	119.959.657	C	T
P2, P3, U	Intron		c.30+188C>A	rs386513902	0.801	-	-	8	119.963.843	G	T
P1, P2, P3, P4, C, U	NSV		c.9C>G	rs2073618	0.357	1.000	0.000	8	119.964.052	G	C
P1, P2, P3, P4, C, U	5'UTR		c.-223C>T	rs2073617	0.408	-	-	8	119.964.283	G	A
C	5'UTR		c.-266C>A	.	-	-	-	8	119.964.326	G	T
P3, P4, C, U	3'UTR		c.*153G>T	rs1053318	0.826	-	-	9	35.056.961	C	A
C, U	Intron		c.2161-241C>T	rs12686362	0.823	-	-	9	35.057.768	G	A
P1, P2, P4	Synonymous		c.1704A>G	rs142577424	0.998	-	-	9	35.059.790	T	C
P1, P2, P3, P4, C, U	Splice region		c.1695+8A>G	rs684562	0.573	-	-	9	35.060.302	T	C
P1, P2, P3, P4, U	Intron		c.1482+52T>C	rs562381	0.306	-	-	9	35.060.746	A	G
P1, P2, P3, P4, C, U	Intron	*VCP*	c.1360-35A>G	rs2258240	0.296	-	-	9	35.060.955	T	C
P1, P3, P4	Intron		c.1194+71A>G	rs2074549	0.859	-	-	9	35.061.503	T	C
P1, P2, P3, P4, C, U	Splice region		c.1082-18_1082-8dupTTGTGTACTGT	rs11272867	0.590	-	-	9	35.061.693	G	GACAGTACACAA
P1, P2, P3, P4, C, U	Splice region		c.811+3G>A	rs514492	0.298	-	-	9	35.062.972	C	T
P1, P2, P4	Intron		c.446-140G>A	rs41274873	0.997	-	-	9	35.065.518	C	T
P1, P3, P4, C, U	Intron		c.129+47G>A	rs10972300	0.844	-	-	9	35.068.201	C	T
P1, P2, U	Upstream		-	rs642347	0.009	-	-	10	13.138.143	A	C
P1, P2, P4	Synonymous	*OPTN*	c.102G>A	rs2234968	0.820	-	-	10	13.151.224	G	A
P1, P2, P3, P4	Intron		c.166+66A>G	rs10906303	0.815	-	-	10	13.151.354	A	G
P1, P4, C, U	Intron		c.167-157A>G	rs60399947	0.861	-	-	10	13.152.117	A	G

P1, P4, C, U	Intron		c.167-150C>T	rs60221241	0.882	-	-	10	13.152.124	C	T
P2	Intron		c.369+29delT	rs398096802	-	-	-	10	13.152.504	GT	G
C	Intron		c.369+190T>G	rs7921853	0.681	-	-	10	13.152.666	T	G
P1, P2, P3, P4, C, U	Splice region		c.553-5C>T	rs2244380	0.205	-	-	10	13.158.262	C	T
P1	Intron		c.780-165T>C	.	-	-	-	10	13.164.220	T	C
P3	Intron		c.780-53T>C	rs765884	0.809	-	-	10	13.164.332	T	C
P1, P2, P3, P4, C, U	Intron		c.882+109A>G	rs489040	0.517	-	-	10	13.164.596	A	G
P2	Intron		c.882+196G>T	rs59554572	0.950	-	-	10	13.164.683	G	T
P1, P2, P3, P4, C, U	NSV	*OPTN*	c.964A>G	rs523747	0.009	1.000	0.000	10	13.166.076	A	G
P1, P2, P3, P4, C, U	Intron		c.1149-86G>T	rs676302	0.199	-	-	10	13.167.860	G	T
P3	Intron		c.1401+389A>C	rs7919563	0.710	-	-	10	13.170.292	A	C
P3, P4	Intron		c.1402-253A>G	rs3740209	0.704	-	-	10	13.173.814	A	G
P2	Intron		c.1532+72G>A	rs77873111	0.978	-	-	10	13.174.269	G	A
P1, P2, P3, P4, C, U	Intron		c.1612+101A>C	rs7086894	0.675	-	-	10	13.175.682	A	C
U	Intron		c.1612+279G>C	rs7068612	0.698	-	-	10	13.175.860	G	C
P1	Intron		c.1613-554G>A	rs11258219	0.789	-	-	10	13.178.191	G	A
P1	Intron		c.1613-544T>A	rs7078784	0.226	-	-	10	13.178.201	T	A
P1, P2, P3, P4, C, U	Intron		c.1613-48C>A	rs10906310	0.734	-	-	10	13.178.697	C	A
P1, P4, U	Intron		c.44+233C>A	rs71430764	0.978	-	-	14	92.980.553	C	A
U	Intron		c.44+10859A>G	rs11622288	0.603	-	-	14	92.991.179	A	G
P1	Intron		c.44+18651C>T	rs6575265	0.296	-	-	14	92.998.971	C	T
P3	Intron		c.45-10541T>C	rs7142824	0.230	-	-	14	93.011.555	T	C
P1	Intron	*RIN3*	c.249+56G>A	rs74072951	0.865	-	-	14	93.022.356	G	A
U	Intron		c.367+214G>A	rs7159046	0.800	-	-	14	93.044.036	G	A
P3	Intron		c.367+862A>G	rs7159925	0.449	-	-	14	93.044.684	A	G
P2	Intron		c.367+8645C>A	rs10136857	0.005	-	-	14	93.052.467	C	A
P3	Intron		c.367+13849G>A	rs4904952	0.005	-	-	14	93.057.671	G	A

P1	Intron		c.368-13818T>C	rs2146499	0.014	-	-	14	93.067.934	T	C
P3	Intron		c.368-3046A>G	rs2181380	0.014	-	-	14	93.078.706	A	G
P2, U	Intron		c.368-99C>T	rs75641131	0.927	-	-	14	93.081.653	C	T
U	Intron		c.440+10873A>G	rs7151638	0.0060	-	-	14	93.092.697	A	G
C	Intron		c.441-5011_441-5008delAGAA	rs142704371	0.850	-	-	14	93.102.571	CAGAA	C
U	Intron		c.533-139_533-134delCAACCT	rs11278705	0.012	-	-	14	93.117.787	ACAACCT	A
P1, P2, P3, P4, C, U	NSV		c.644A>G	rs3829947	0.566	-	0.000	14	93.118.038	A	G
U	Synonymous		c.804C>T	rs3814830	0.735	-	-	14	93.118.198	C	T
C, U	NSV		c.1274C>T	rs3742717	0.725	0.270	0.014	14	93.118.668	C	T
P1, P3, P4, C, U	Synonymous		c.1275G>A	rs3742716	0.717	-	-	14	93.118.669	G	A
C	Synonymous	*RIN3*	c.2013C>T	rs3818321	0.860	-	-	14	93.119.407	C	T
P1, C	Intron		c.2027-2929C>T	rs8013795	0.011	-	-	14	93.122.577	C	T
P1	Intron		c.2027-224G>T	rs12885166	0.714	-	-	14	93.125.282	G	T
P2, C	Intron		c.2027-91G>C	rs2295991	0.686	-	-	14	93.125.415	G	C
P2	Intron		c.2336-326A>G	rs2273926	0.720	-	-	14	93.142.494	A	G
P2	Intron		c.2336-286T>C	rs2273925	0.746	-	-	14	93.142.534	T	C
P2, C, U	Intron		c.2336-183G>A	rs2273924	0.719	-	-	14	93.142.637	G	A
P2, C, U	Intron		c.2336-175T>C	rs2273923	0.720	-	-	14	93.142.645	T	C
P1, P4	Intron		c.2336-169G>A	rs733447	0.790	-	-	14	93.142.651	G	A
P2	NSV		c.2377T>C	rs147042536	0.997	0.000	0.999	14	93.142.861	T	C
U	Intron		c.2632-859G>A	rs61994063	0.763	-	-	14	93.153.412	G	A
P1, P4, C, U	NFD		c.2899_2901delGGC	rs71698059	0.331	-	-	14	93.154.537	TGGC	T
P2	Intron		c.602+206G>A	rs3784562	0.410	-	-	15	74.291.023	G	A
P1, P2, C	Intron		c.1184-155T>C	rs2277599	0.440	-	-	15	74.317.043	T	C
C	Intron		c.1254+603C>G	rs12902857	0.437	-	-	15	74.317.871	C	G
P3	Intron	*PML*	c.1398+126A>G	rs2304716	0.401	-	-	15	74.325.182	A	G
P1, P3	Intron		c.1658-108G>A	rs2304718	0.506	-	-	15	74.326.711	G	A
P1, P2, P3, C	Intron		c.1710+1245A>G	rs743580	0.451	-	-	15	74.328.116	A	G

C	Intron		c.1710+1270G>T	rs743581	0.613	-	-	15	74.328.141	G	T
P1, P2, P3	Intron		c.1710+1335G>C	rs743582	0.882	-	-	15	74.328.206	G	C
P1, P3	Intron	*PML*	c.1710+1705A>G	rs9479	0.466	-	-	15	74.328.576	A	G
P1, P2, P3, P4	Intron		c.1862-168C>T	rs8032123	0.059	-	-	15	74.336.394	C	T
P1, P2, P3, P4	NSV		c.1933T>C	rs5742915	0.773	0.900	0.001	15	74.336.633	T	C
P3, U	Intron	*GOLGA6A*	c.1593+81A>G	rs79015760	0.802	-	-	15	74.364.478	T	C
P1, P2	Intron		c.76-194G>C	rs11152342	0.697	-	-	18	60.015.207	G	C
P2	Intron		c.284-120T>G	rs3826620	0.339	-	-	18	60.021.504	T	G
P3, P4, U	NSV		c.421C>T	rs35211496	0.920	0.130	0.762	18	60.021.761	C	T
P2	Intron		c.428-22delT	.	-	-	-	18	60.025.458	GT	G
U	Intron		c.428-22_428-21delTT	rs71160827	-	-	-	18	60.025.458	GTT	G
P3, P4	Intron		c.522-183T>A	rs6567270	0.528	-	-	18	60.027.005	T	A
P1, P3, P4, C, U	Intron		c.522-39T>A	rs6567271	0.395	-	-	18	60.027.149	T	A
P1, P3, P4, C, U	Intron		c.522-17C>T	rs6567272	0.395	-	-	18	60.027.171	C	T
P1, P3, P4, C, U	NSV	*TNFRSF11A*	c.575C>T	rs1805034	0.395	1.000	0.000	18	60.027.241	C	T
P1, P3, P4, C, U	Intron		c.616+79G>A	rs9653064	0.433	-	-	18	60.027.361	G	A
U	Intron		c.617-258A>C	rs8083511	0.657	-	-	18	60.028.655	A	C
P1, P3, C, U	Intron		c.617-151G>A	rs8099222	0.769	-	-	18	60.028.762	G	A
C	Intron		c.730+212G>C	rs7239667	0.547	-	-	18	60.029.238	G	C
P1, P2, P3, P4, C, U	Synonymous		c.933A>G	rs8092336	0.022	-	-	18	60.036.083	A	G
C	NSV		c.1519G>A	rs61751992	0.996	0.040	0.276	18	60.036.669	G	A
C	Intron		c.1568-399A>T	rs9956633	0.015	-	-	18	60.051.585	A	T
P1, C	Intron		c.1568-43C>T	rs56231704	0.876	-	-	18	60.051.941	C	T

CDS change: Coding DNA sequence change; Chr.: Chromosome; Ref.: Reference; Obs.: Observed; NSV: Non-synonymous Variants; NFD: Non-Frameshift Deletion.

Note – Individual: P1 - III.2-080004; P2 - IV.9-090044; P3 - III.4-080001; P4 - III.6-080005; C - IV.1-080002; U - IV.3-090095.

Individual IV.1-080002 is a control, the remaining (III.4-080001, III.2-080004, III.6-080005, IV.9-090044 and IV.3-090095) are PDB affected, for model 1.

Individuals IV.1-080002 and IV.3-090095 are controls, the remaining (III.4-080001, III.2-080004, III.6-080005 and IV.9-090044) are PDB affected, for model 2.